Semantic Relations: Discovery and Applications

Tutorial

Roxana Girju
University of Illinois at Urbana-Champaign

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

SEMANTIC FRONTIERS

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Applications of semantic relations
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

2

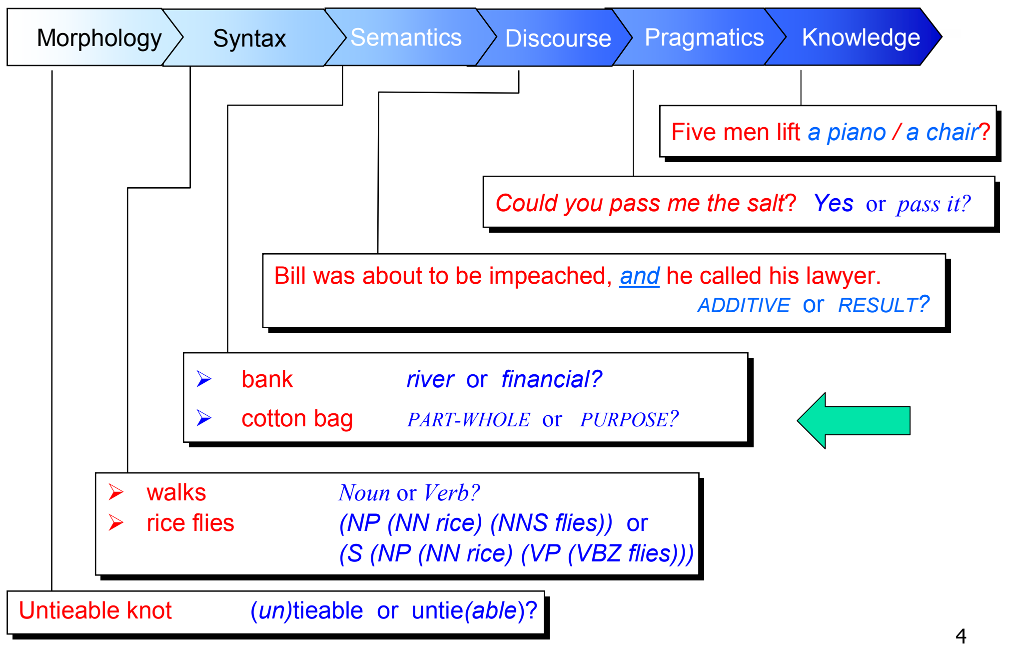
The Problem of Knowledge Discovery

Definitions:

- *Information, Data* - a collection of facts from which inferences can be drawn
- *Knowledge, cognition* - the psychological result of perception and learning and reasoning.
- *Knowledge Discovery from Text* - is the process of extracting both explicit and implicit knowledge from unstructured data.

3

Levels of Language Analysis - Computational challenges



4

Semantic relations

- are underlying relations between two concepts expressed by words or phrases

- play an essential role in lexical semantics
- applications:
  - Lexico-semantic knowledge bases,
  - Question Answering, Text Summarization, Textual Entailment, Text-to-Image Generation, etc.
- Examples:
  - HYPERNYMY (IS-A),
  - MERONYMY (PART-WHOLE),
  - CAUSE - EFFECT, etc.

5

Knowledge Discovery

- *Knowledge Discovery* is the extraction of non-trivial, useful information from data.
  - Why Discovery?
    - Semantics (meaning of words/phrases) is often implicit

- How can we discovery semantic relations?
  - Semantic parsing = the process of mapping a natural-language sentence into a formal representation of its meaning.
  - A deeper semantic analysis provides a representation of the sentence in formal language which supports automated reasoning.

6

Discovery of Semantic Relations (1)

Unstructured Information

Structured Knowledge

Web
Documents
News
Digital library

Semantic Parser

- KB
- Semantically tagged text
  - Concepts
  - Semantic elations
  - Links between multiple docs

The following examples illustrate the problem of semantic relation discovery from text presented in this tutorial.

7

Discovery of Semantic Relations (2)

Example 1:

[*Saturday's snowfall*]TEMP topped [*a record* in *Hartford, Connecticut*]LOC with [*the total of 12.5 inches*]MEASURE, [the *weather service*]TOPIC said. The storm claimed its fatality Thursday when [a *car* driven by a [*college student*]PART-WHOLE]THEME skidded on [*an interstate overpass*]LOC in [*the mountains of Virginia*]LOC/PART-WHOLE and hit [*a concrete barrier*]PART-WHOLE, police said.

(www.cnn.com – "Record-setting Northeast snowstorm winding down", December 7, 2003)

TEMP (Saturday, snowfall)
LOC (Hartford Connecticut, record)
MEASURE (total, 12.5 inch)
TOPIC (weather, service)
PART-WHOLE (student, college)
THEME (car, driven by a college student)

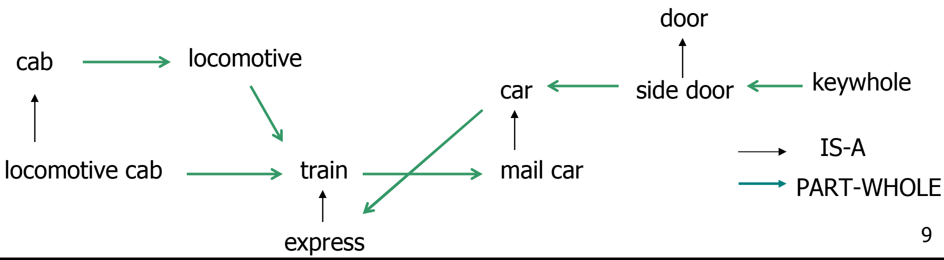
LOC (interstate, overpass)
LOC (mountains, Virginia)
PART-WHOLE/LOC (mountains, Virginia)
PART-WHOLE (concrete, barrier)

8

Discovery of Semantic Relations (3)

Example 2:

The car's mail messenger is busy at work in [the *mail car*]PART-WHOLE as the train moves along. Through the open [[*side door*]PART-WHOLE of the *car*]PART-WHOLE, moving scenery can be seen. The worker is alarmed when he hears an unusual sound. He peeks through [the *door*'s *keyhole*]PART-WHOLE leading to the tender and [*locomotive cab*]PART-WHOLE and sees the two bandits trying to break through [the [*express car*]PART-WHOLE *door*]PART-WHOLE.



Discovery of Semantic Relations (4)

Example 3:

Colleagues today recall [with some humor] [how] meetings would crawl into the early morning hours as Mr. Dinkins would [quietly] march his staff out of board meetings and into his private office to discuss, [en masse], certain controversial proposals the way he knows [best].

MANNER (with some humor, recall)
MANNER (how, crawl)
MANNER (quietly, march)
MANNER (en masse, discuss)
MANNER (the way he knows, discuss)
MANNER (best, knows)

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Application of KDT
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

11

Motivation (1)

Semantic relation discovery has both theoretical and practical implications:

- In the past few years it has received considerable attention; E.g.:
  - Workshop on Multilingual Expressions (COLING/ACL 2006, 2004, 2003)
  - Workshop on Computational Lexical Semantics (ACL 2004)
  - Tutorial on Knowledge Discovery from Text (ACL 2003)
  - Shared task on Semantic Role Labeling (CoNLL 2004, 2005, 2008; SENSEVAL 2004)
  - SEMEVAL 2007, Task 4 (ACL 2007): Semantic Relations between Nominals
- It has a large number of applications:
  - Question Answering
  - Textual Entailment
  - Text-to-Image Generation
  - Etc.

12

Motivation (2)

Semantic relation discovery has both theoretical and practical implications:

- It has been part of major international projects related on knowledge discovery:
  - ACE (Automatic Content Extraction)
    - http://www.itl.nist.gov/iad/894.01/tests/ace/
  - DARPA EELD (Evidence Extraction and Link Discovery)
    - http://w2.eff.org/Privacy/TIA/eeld.php
  - ARDA-AQUAINT (Question Answering for Intelligence)
  - ARDA NIMD (Novel Intelligence from Massive Data)
  - Global WordNet
    - http://www.globalwordnet.org/
  - Etc.

13

Motivation (3)

Knowledge intensive applications

E.g. ***Question Answering***

Q: What does the [BMW company]IS-A produce?
A: "[BMW cars]MAKE-PRODUCE are sold .."

Q: Where have nuclear incidents occurred?
A: "The [(Three Mile Island) (nuclear incident)]LOC caused a DOE policy crisis.."

Q: What causes malaria?
A: "..to protect themselves and others from being bitten by [malaria mosquitoes]CAUSE.."

14

Motivation (4)

Knowledge intensive applications

Q: What does the AH-64A Apache helicopter consist of?

(Girju et al., 2003)

A:
AH-64A Apache helicopter
  Hellfire air to surface missille
      millimeter wave seeker
    70mm Folding Fin Aerial rocket
    30mm Cannon camera
    armaments
    General Electric 1700-GE engine
    4-rail launchers
    four-bladed main rotor
    anti-tank laser guided missile
    Longbow millimetre wave fire control radar
      integrated radar frequency interferometer
    rotating turret
    tandem cockpit
      Kevlar seats

(Defense Industries: www.army-technology.com)

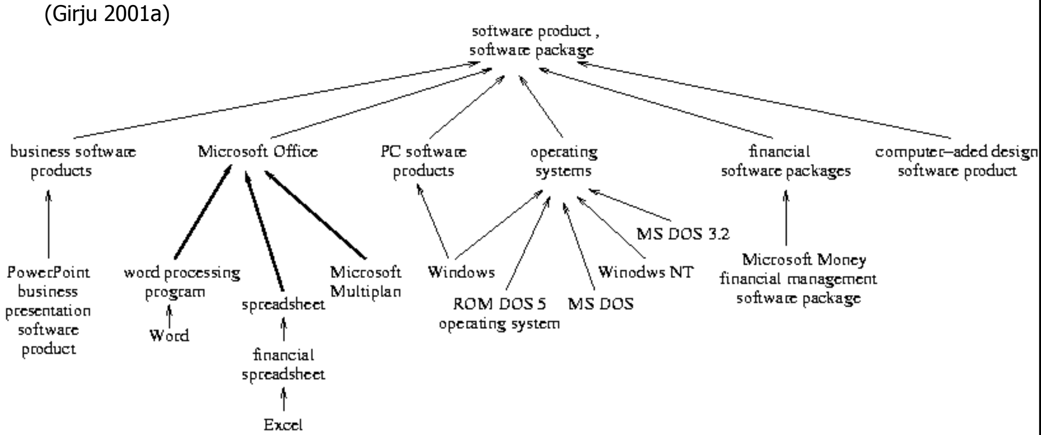
15

Motivation (5)

Knowledge intensive applications

Q: What *software products* does Microsoft sell?

(Girju 2001a)



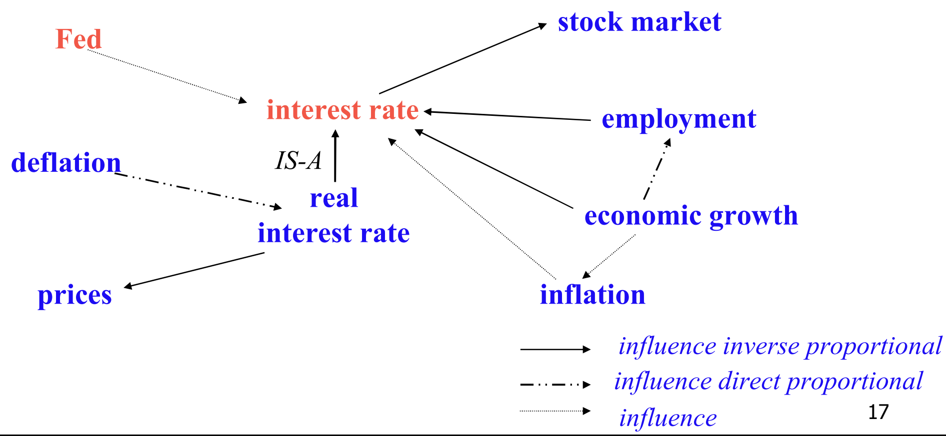
16

Motivation (6)

Knowledge intensive applications

Q: Will the Fed change interest rate at their next meeting?

(Girju 2001b)



17

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Applications of semantic relations
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

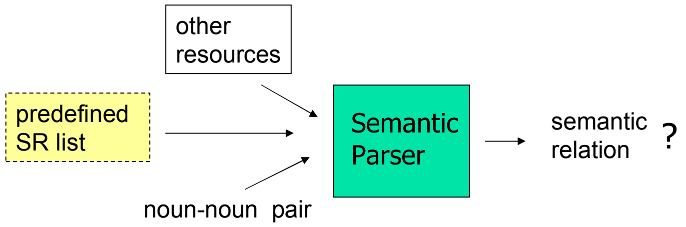
18

Basic Approaches (1)

The task of semantic relation discovery:

Given a pair of nouns n1 – n2, determine the pair's meaning.

Q: How is the meaning expressed?



19

Basic Approaches (2)

Currently, there are two main approaches in Computational Linguistics:

1. Labeling:

Task: Given a noun – noun instance, label it with the underlined semantic relation

Requirements: A predefined list of semantic relations

Example: summer vacation → TEMPORAL

2. Paraphrasing:

Task: Given a noun – noun instance, find a paraphrase that preserves the meaning in context

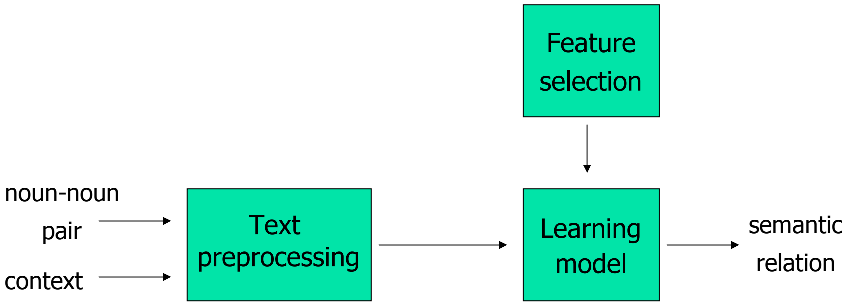
Example: summer vacation → vacation during summer

Q: Which approach is better?

20

Basic Approaches (3)

Semantic parsers:



21

Basic Approaches (4)

- Text processing:
  - Tokenizer
  - Part-of-speech tagger
  - Syntactic parser
  - Word sense disambiguation
  - Named entity recognition
  - Etc.
- Feature selection:
  - Determines the set of characteristics (constraints) of the nouns and/or context to include in the classifier in order to differentiate among semantic relations
- Classifier:
  - Classifies various input instances into corresponding semantic relations; Usually a machine learning model

22

Basic Approaches (5)

Tokenizer - brakes a document into lexical entities called tokens

E.g.: U.S.A. U, ., S, ., A, .

- Tokens are:
  - Alphanumerical characters and strings
  - Numbers
  - Genitives ', 's
  - SGML tags
  - Common multi-character separators

23

Basic Approaches (6)

Part of speech (POS) tagger - labels each word with its corresponding part of speech in context

E.g.: Mary/NNP has/VBZ a/DT cat/NN ./.

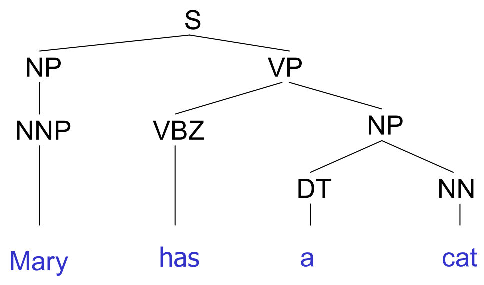
DT – determiner
NN – common noun
NNP – proper noun
VBZ – verb at present tense

24

Basic Approaches (7)

Syntactic parser - groups words into phrases

E.g.: Mary has a cat



Basic Approaches (8)

Word sense disambiguation - identifies word senses in context

E.g.: The *bank#2* is open until 7pm.
They pulled the canoe up on the *bank#1*.

WordNet:

bank#1 – river bank
bank#2 – financial institution

Basic Approaches (9)

Named-Entity Recognizer

NE Recognizer identifies named entities such as:

- *Organizations:* Michigan State university, Dallas Cowboys, U.S. Navy,
- *Locations:* "Dallas, TX", "Frascati, Italia", Lake Tahoe, Bay area, African Coast.
- *Persons:* Mr. Smith, Deputy Smith
- *Addresses:* 750 Trail Ridge
- *Other names:* Austin

27

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Applications of semantic relations
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

28

Semantic relation discovery – The Challenges (1)

- Natural Language has many ambiguities.
- Open domain text processing is very difficult
- There is no general agreement on basic issues
  - What is a concept
  - What is a context
  - How to represent text knowledge

29

Semantic relation discovery – The Challenges (2)

- Semantic relations are encoded at various lexico-syntactic levels:
  - e.g., N1 N2 (tea cup), N2 prep. N1 (cup of tea), N1's N2 (*tea's cup);
- The compounding process (N N) is highly productive, but not totally constrained:
  - "war man" is not a man who hates the war (Zimmer 1971, cf. Downing 1977).
- Semantic relations are usually implicit; Examples:
  - spoon handle (whole-part);
  - bread knife (functional);
- Semantic relation discovery may be knowledge intensive:
  - Eg: GM car

30

Semantic relation discovery – The Challenges (3)

- There can be many possible relations between a given pair of noun constituents:
  - firewall: wall to keep out fire; network security software;
  - the girl's shoes: the ones she owns, she dreams about, she made herself, etc.
  - Texas city: Location / Part-Whole
- Interpretation can be highly context-dependent:
  - apple juice seat: seat with apple juice on the table in front of it (Downing 1077).
- There is no well defined set of semantic relations:
  - Very abstract: of, by, with, etc.; (Lauer 1995)
  - Very specific: dissolved in, etc.; (Finin 1980)

Outline

Lists of Semantic Relations: Approaches in Linguistics (1)

Most of the research on relationships between nouns and modifiers deals with noun compounds, but these can also hold between nouns and adjective premodifiers/postmodifiers prepositional phrases (complex nominals).

What are Noun Compounds?

- Definition:
  - typically binary, but can also be longer than two words; commonly written as a sequence of words, single words, or single words linked by hyphens;

- Frequently used in:
  - technical writing and newswire text (McDonald 1982);
  - fictional prose (Leonard 1984);
  - spoken language (Liberman & Sproat 1992);
  - medical domain (Rosario and Hearst 2001, 2002);

Lists of Semantic Relations: Approaches in Linguistics (2)

Different terminology in Theoretical Linguistics and NLP:

▪ substantive compound:	Jespersen 1954;
▪ nominal compound:	Lees 1960, 1970;
▪ complex nominal (CN):	Levi 1978;
▪ noun + noun compound:	Downing 1977;
▪ nominal compound:	Finin 1980;
▪ noun sequence:	Leonard 1984; Vanderwende 1994, 1995;

Lists of Semantic Relations: Approaches in Linguistics (3)

- Why should we care about complex nominals?
  - >1% of words in the BNC participate in CNs (Tanaka and Baldwin 2003)
  - 70% are singletons (Lapata and Lascarides 2003)
  - 35% of CNs involve nominalizations (Grover et al. 2005, Nicholson 2005)

35

Lists of Semantic Relations: Approaches in Linguistics (4)

Types of Complex Nominals:

1. Classification based on whether the meaning of the CN can be derived from the meaning of the constituents:

- compositional CNs:
  - have a literal interpretation;
  - E.g.: spoon handle = handle is part of the spoon
- non-compositional (lexicalized) CNs:
  - the meaning is a matter of convention;
  - cannot be synthesized from the meanings of their space-limited components;
  - neither noun plays any role in the semantic interpretation (Levi 1978) and for which no literal interpretation is appropriate;
  - E.g.: *soap opera*;

36

Lists of Semantic Relations: Approaches in Linguistics (5)

2. Classification based on the type of the head noun:

- Verbal compounds (synthetic compounds):
  - the head noun is derived from the underlying predicate (a verb, but possibly an adjective too);
  - the modifier fulfills an argument function, or has a thematic role in relation to the verbal head;
  - E.g.: music lover
- Non-verbal compounds (primary/root compounds):
  - the head noun is not derived from the underlying predicate, or it is derived from a verb but the modifier is not one of its arguments;
  - E.g.: spoon handle

37

Lists of Semantic Relations: Approaches in Linguistics (6)

Summary of previous work:

(Jespersen 1954, Lees 1960, 1970, Downing 1977, Levi 1978, Warren 1978, Bauer 1979, Selkirk 1982, Lieber 1983, Isabelle 1984, Jensen 1990, etc.)

- CN studied from various perspectives, but mainly focusing on the semantic aspect;
- The CN interpretation problem was tackled by providing a classification schema (Jespersen, Lees, and Levi):
  - Focus on lexicalized compounds (e.g.: swan song)
  - Differences in the theoretical frameworks used
  - Similarities: all require prior knowledge of the meaning conveyed by the CN
  - No agreement among linguists on a single such classification

38

Lists of Semantic Relations: Approaches in Linguistics (7)

The syntactic approach to N – N interpretation:
(Lees 1960, 1970; Levi 1978; Selkirk 1982; Grimshaw 1991)

- from the Generative Semantics perspective
  - it was assumed that the interpretation of compounds was available because the examples were derived from underlying relative clauses that had the same meanings.
  - E.g.: *honey bee*, expressing the relation MAKE, was taken to be derived from a headed relative *a bee that makes honey*.

39

Lists of Semantic Relations: Approaches in Linguistics (8)

The syntactic approach to N – N interpretation:
(Lees 1960, 1970; Levi 1978; Selkirk 1982; Grimshaw 1991)

- Interpretation based on grammatical criteria using a transformational approach
- Semantic content of a noun compound is characterized by means of a sentential paraphrase;
- Finite number of syntactic and semantic relationships which underlie the various classes of NCs;
- Generation of these relationships is fully productive;

40

Lists of Semantic Relations: Approaches in Linguistics (9)

(Levi 1978):

- focus: syntactic and semantic properties of NCs
- complex nominals (includes nominal nonpredicating adjectives as possible modifiers);
- NCs are generated according to a set of transformations from underlying relative clauses or complement structures. Two syntactic processes are used:
  - predicate nominalization;
  - predicate deletion;

41

Lists of Semantic Relations: Approaches in Linguistics (10)

(Levi 1978):

- Two syntactic processes are used:
  - predicate nominalization:
    - those involving nominalizations, i.e., compounds whose heads are nouns derived from a verb, and whose modifiers are interpreted as arguments of the related verb
    - E.g.: "x such that x plans cities" => city planner;
  - predicate deletion:
    - List of relations: ***cause, have, make, use, be, in, for, from, about***
    - E.g.: "field mouse" derived from "a mouse which is in the field" ("in" deletion);
    - Deleted predicates represent the only semantic relations which can underlie NCs not formed through predicate nominalization;

42

Lists of Semantic Relations: Approaches in Linguistics (11)

(Levi 1978):

- Disadvantages:
  - uses a limited list of predicates that are primitive semantically and not sufficient to disambiguate various NCs:
  - E.g., headache pills (FOR), fertility pills (FOR); (cf. Levi 1978)

43

Lists of Semantic Relations: Approaches in Linguistics (12)

The pragmatic approach to N-N interpretation:
(Downing 1978)

- psycho-linguistic approach;
- focuses on statistical knowledge to interpret novel pairings;
- relevant from the point of view of production, rather than interpretation
- criticized previous approaches on that the interpretation of CNs involve pragmatic knowledge;
- covers only N-N compounds;
  - E.g.: *apple juice seat* – "a seat in front of which an apple juice [is] placed" (Downing, 1977 page 818) – which can only be interpreted in the current discourse context.

44

Outline

Lists of Semantic Relations: Approaches in NLP (1)

- Have followed mostly the proposals made in theoretical linguistics;
- Rely on sets of semantic relations of various sizes and at different abstraction levels:
  - 8 prepositions: of, for, in, at, on, from, with, about (Lauer, 1995)
  - Thousands of specific verbs: dissolved in, etc. (Finin 1980)
  - No general agreement

Lists of Semantic Relations: Approaches in NLP (2)

- State-of-the-art lists of semantic relations used in the literature:

1) A list of 8 prepositions (Lauer 1995);
2) A two-level taxonomy of semantic relations (Barker and Szpakowicz 1998; Nastase and Szpakowicz 2003);
3) A list of 22 semantic relations (Moldovan & Girju 2004; Girju 2006);
4) A list of 7 semantic relations (SemEval 2007 – Task 4);

The last three sets overlap considerably

47

Lists of Semantic Relations: Approaches in NLP (3)

A two-level taxonomy of semantic relations

(Barker, K., and Szpakowicz, S. 1998; Nastase, V., and Szpakowicz, S. 2003);

- Some examples (H denotes the head of a base NP, M denotes the modifier):

Semantic relations		Example
Causal	Cause	flu virus
	Effect	exam anxiety
	Purpose	concert hall
Participant	Agent	student protest
	Object	metal separator
	Beneficiary	student discount
Spatial	Direction	outgoing mail
	Location	home town
	Location at	desert storm

48

Lists of Semantic Relations: Approaches in NLP (4)

A two-level taxonomy of semantic relations

(Barker, K., and Szpakowicz, S. 1998; Nastase, V., and Szpakowicz, S. 2003);

Semantic relations		Example
Temporal	Frequency	weekly game
	Time at	morning coffee
	Time through	2-hour trip
Quality	Manner	stylish writing
	Material	brick rock
	Measure	heavy rock

49

Lists of Semantic Relations: Approaches in NLP (5)

A list of 22 semantic relations

(Moldovan & Girju 2004; Girju 2006);

Semantic Relation	Definition/ Example
HYPERNYMY (IS-A)	an entity/event/state is a subclass of another; (*daisy flower; large company, such as Microsoft*)
PART-WHOLE (MERONYMY)	an entity/event/state is a part of another entity/event/state; (*door knob; the door of the car*);
CAUSE	an event/state makes another event/state to take place; (*malaria mosquitos; "death by hunger"; "The earthquake generated a big Tsunami"*);

50

Lists of Semantic Relations: Approaches in NLP (6)

POSSESSION	an animate entity possesses (owns) another entity; (*family estate; the girl has a new car.*)
KINSHIP	an animated entity related by blood, marriage, adoption or strong affinity to another animated entity; (*boy's sister; Mary has a daughter*)
MAKE/PRODUCE	an animated entity creates or manufactures another entity; (*honey bees; GM makes cars*)
INSTRUMENT	an entity used in an event as instrument; (*pump drainage; He broke the box with a hammer.*)

51

Lists of Semantic Relations: Approaches in NLP (7)

TEMPORAL	time associated with an event; (*5-O' clock tea; the store opens at 9 am*)
LOCATION/ SPACE	spacial relation between two entities or between an event and an entity; (*field mouse; I left the keys in the car*)
PURPOSE	a state/activity intended to result from another state/event; (*migraine drug; He was quiet in order not to disturb her.*)
SOURCE/FROM	place where an entity comes from; (*olive oil*)

52

Lists of Semantic Relations: Approaches in NLP (8)

EXPERIENCER	an animated entity experiencing a state/feeling; (*desire for chocolate; Mary's fear*.)
TOPIC	an object specializing another object; (*they argued about politics*)
MANNER	a way in which an event is performed or takes place; (*hard-working immigrants; performance with passion*)
MEANS	the means by which an event is performed or takes place; (*bus service; I go to school by bus.*)

53

Lists of Semantic Relations: Approaches in NLP (9)

AGENT	the doer of an action; (the investigation of the police)
THEME	the entity acted upon in an action/event (music lover)
PROPERTY	characteristic or quality of an entity/event/state; (*red rose; the juice has a funny color.*)
BENEFICIARY	an animated entity that benefits from the state resulting from an event; (*customer service; I wrote Mary a letter*.)

54

Lists of Semantic Relations: Approaches in NLP (10)

MEASURE	an entity expressing quantity of another entity/event; (*70-km distance; The jacket costs $60; a cup of sugar*)
TYPE	a word/concept is a type of word/concept; (*member state; framework law*)
DEPICTION-DEPICTED	an entity is represented in another; (*the picture of the girl*)

55

Outline

56

Architectures of Semantic Parsers (1)

Currently, there are two main approaches in Computational Linguistics:

1. Labeling:

Task: Given a noun – noun instance, label it with the underlined semantic relation

Requirements: A predefined list of semantic relations

Example: summer vacation → TEMPORAL

2. Paraphrasing:

Task: Given a noun – noun instance, find a paraphrase that preserves the meaning in context

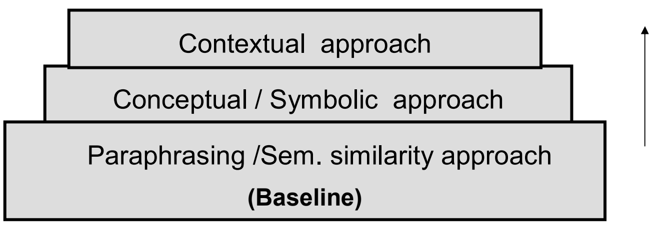
Example: summer vacation → vacation during summer

Q: Which approach is better?

57

Architectures of Semantic Parsers (2)

Taxonomy of basic semantic parsers:



58

Architectures of Semantic Parsers (3)

Criteria	Approaches	
	Paraphrasing	**Labeling**
Context	Out-of-context	Mostly in-context (with some exceptions)
Syntactic level (pattern-based)	Pattern-based; Usually "N N", " N P N", "N vb N"	Mostly pattern-based; but not recently (e.g., SemEval 2007)
Knowledge (resource dependent)	Knowledge-poor; Mostly based on frequencies on large corpora;	Knowledge-intensive (mostly; usually based on WordNet)
Learning method	Unsupervised or Weakly supervised	Supervised

59

Architectures of Semantic Parsers (4)

Earlier work on Noun – Noun interpretation:

(Finin 1980, 1986; McDonalds 1982; Sparck Jones 1983; Leonard 1984; Lehnert 1988; Riloff 1989; Vanderwende 1994, 1995; Lauer 1995, 1996; Fabre 1996; Fabre and Sebillot 1995; ter Stal 1996; Barker 1998; Lapata 2000; Rosario & Hearst 2001; Rosario, Hearst & Fillmore 2002, etc.)

- Approaches:
  - often based on the analysis of the semantics of the individual nouns; this assumes some existence of a dictionary of semantic information;
  - classification:
    - Symbolic, MRD-based systems (most of them domain specific),
    - Statistical (unsupervised models);

60

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Applications of semantic relations
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

61

Paraphrasing Systems (1)

Statistical / Unsupervised Approaches:

(Lauer 1995a, 1996):

- Uses as semantic relations paraphrases, clauses or prepositional phrases to illustrate relationships in noun-noun compounds
- focuses only on nouns acting as single nouns;
- 8 relations: of, for, in, at, on, from, with, about;
- Assigns probabilities to each of the different possible paraphrases of a compound based on the probability distribution of the relations in which the modifier and header are likely to participate;

62

Paraphrasing Systems (2)

- Uses a large corpus to compute frequencies of prepositions to estimate probabilities;
- Maps words in CNs into categories in Roget's Thesaurus and finds probabilities of occurrence of certain NCs and their paraphrases;
- No automatic process in finding the best level of generalization
- His approach only applies to non-verbal noun compounds, non-copulative CNs.

- Interprets "*a b*" as "*b* <prep> *a*", where <prep> is one of: of, for, in, at, on, from, with, about.
  - state laws → "laws *of the* state"
  - baby chair→ "chair *for* babies"
  - reactor waste → "waste *from* a reactor"

63

Paraphrasing Systems (3)

Predicting paraphrases:

- When predicting which preposition to use in the paraphrase, it is a simple case of choosing the most probable.

$$p^* = \arg\max_p P(p \mid n_1, n_2) \qquad (1)$$

After some assumptions regarding independence and uniformity, and applying Bayes' theorem, this simplifies to (t1 and t2 are concepts in Roget's thesaurus):

$$p^* = \arg\max_p \sum_{\substack{t_1 \in cats(n_1) \\ t_2 \in cats(n_2)}} P(t_1 \mid p) P(t_2 \mid p) \qquad (2)$$

64

Paraphrasing Systems (4)

Experiments:

- Lauer tested the model on 282 compounds that he selected randomly from Grolier's encyclopedia and annotated with their paraphrasing prepositions.
- The preposition *of* accounted for 33% of the paraphrases in this data set.
- The concept based model (see (2)) achieved an accuracy of 28% on this test set, whereas its lexicalized version reached an accuracy of 40%.

65

Paraphrasing Systems (5)

Results:

- Overall, the results are abysmal, only barely reaching significance above the baseline of always guessing *of* (the most common paraphrase)
- Word based counts tend to perform marginally better than class smoothed counts
- Restricting guesses only the most common results can significantly increase accuracy, but at the cost of never guessing the less frequent relations.

66

Paraphrasing Systems (6)

Statistical / Unsupervised Approaches:

(Lapata & Keller 2005):

- follow and improve over Lauer's approach
- have analyzed the effect of using Internet search engine result counts for estimating probabilities, instead of a standard corpus

67

Paraphrasing Systems (7)

- On the noun compound interpretation problem, they compared:
  - Web-based n-grams (unsupervised)
  - BNC (smaller corpus)-based n-grams
- In all cases, the Web-based n-grams were:
  - The same as or better than BNC-based n-grams
- Thus, they propose using the Web as a baseline against which noun – noun interpretation algorithms should be compared.

68

Paraphrasing Systems (8)

Computing Web n-grams:

- Find # of hits for each term via Altavista / Google:
  - This gives document frequency, not term frequency
  - Smooth 0 counts to 0.5
  - All terms lower case

69

Paraphrasing Systems (9)

Noun compound interpretation

- Determine the semantic relation between nouns; E.g.:
  - war story -> story **about** war
  - pet spray -> spray **for** pet
- Method:
  - Look for prepositions that tend to indicate the relation
  - Used inflected queries (inserted determiners before nouns):
    - Story/stories about the/a/0 war/wars

70

Paraphrasing Systems (10)

Noun compound interpretation

- Results:
  - Best scores obtained for f(n1, p, n2)
  - Significantly outperforms the baseline and Lauer's model
  - Show the Web as a corpus is much better as BNC
  - So: a baseline that must be beat in order to declare a new interpretation model to be useful.

Paraphrasing Systems (11)

Model	Altavista	BNC
$f(n_1,p)f(p,n_2)$	50.71	27.85#
$f(n_1,p,n_2)$	55.71#*	11.42
$f(n_1,p)f(p,n_2)/f(p)$	47.14	26.42
$f(n_1,p,n_2)/f(p)$	55.00	10.71

Table 10: Performance of Altavista counts and BNC counts for compound interpretation (data from Lauer 1995)

Model	Accuracy
Best BNC	27.85†‡
Lauer (1995): concept-based	28.00
Baseline	33.00
Lauer (1995): word-based	40.00
Best Altavista	55.71†‡

Table 11: Performance comparison with the literature for compound interpretation

Paraphrasing Systems (12)

Prepositional paraphrases:

- Pros
  - Small list of classes
  - Easily identified in corpus texts
  - Commonly used

- Cons
  - Does not always apply. (E.g.: $100 scarf, flu virus)
  - Very shallow representation of semantics
  - Certain nouns present various lexical preferences for various prepositions, which can skew empirical results
  - Some relations can be expressed by multiple prepositions

73

Paraphrasing Systems (13)

Other observations on Lapata & Keller's unsupervised model:

- We replicated their model on a set of noun compounds from literary work and Europarl (European Parliament sessions)
- We manually checked the first five entries of the pages returned by Google for each most frequent preposition paraphrase

N-N compound test set	Ambiguity of noun constituents	Accuracy [%]
Set #1	One POS, one WN sense	35.28
Set #2	Multiple POS, one WN senses	31.22
Set #3	One POS, multiple WN senses	50.63
Set #4	Multiple POS, multiples WN senses	43.25

74

Paraphrasing Systems (14)

Results:

- For sets #2 and #4, the model introduces a number of false positives:
  - E.g.: baby cry generated ".. it will make moms cry with the baby"

- 30% of the noun compounds in sets #3 and #4 had at least two possible readings:
  - E.g.: paper bag → bag for papers (Purpose)
    paper bag → bag of paper (Material/Part-Whole)

75

Outline

76

Symbolic / Conceptual-based Systems (1)

Symbolic systems:

(Finin 1980, 1986; McDonald 1986; Ter Stal 1996; Vanderwende 1994; Barker 1998; Rosario & Hearst 2001; Rosario, Hearst and Fillmore 2002)

(Finin 1980, 1986; McDonald 1986):

- Based on ad-hoc hand-coded dictionaries; concept dependent;
- Noun compounds: aircraft engine (Part-Whole);
- Systems:
  - Input: two-word NCs;
  - Output: semantic class or multiple interpretations of a compound with appropriate score attached;
  - Representation based on individual nouns mapped onto concepts characterized by a set of roles and slots and arranged in an abstraction hierarchy;
  - Used to support semantic interpretation rules;
- Rely on lexical information that cannot be acquired automatically => unsuitable for unrestricted text.

77

Symbolic / Conceptual-based Systems (2)

Classification schema: (Finin 1980)

- Lexical interpreter:
  - maps incoming surface words into one or more underlying concepts;
- Concept modifier:
  - produces a set of scored possible interpretations between a given head concept and a potential modifying concept;
- Modifier parser:
  - compares and combines the local decisions made by the other two components to produce a strong interpretation;

78

Symbolic / Conceptual-based Systems (3)

Classes of interpretation rules:

- Idiomatic rules (relationship independent of the constituents); e.g., hanger queen;
- Productive rules (general patterns which can produce many instantiations); characterized by semantic relationships;
  - E.g.: rule for dissolved in:
    - Modifier: chemical compound;
    - Modified: liquid, preferably water;
- Structural rules (characterized by structural relationships between modifying and modified concepts);

79

Symbolic / Conceptual-based Systems (4)

Symbolic systems:

(Vanderwende 1994):

- SENS system designed for analyzing CNs in unrestricted text;
- Attempts to avoid hand-coding required by previous attempts;
- Extracts automatically semantic features of nouns from on-line dictionary definitions;
- Algorithm:
  - Input: two noun NCs with no context considered;
  - Output: an ordered list of interpretations
  - uses a set of general rules with associated weights and a general procedure for matching words. Checks how closely related two words are.

80

Symbolic / Conceptual-based Systems (5)

- NC interpretation:
  - classes studied previously in theoretical linguistics (Downing 1977; Jespersen 1954; Lees 1960; Levi 1978)
  - The classification schema has been formulated as wh-questions;
- No WSD;

81

Symbolic / Conceptual-based Systems (6)

Question	Relation Name	Example
Who/what?	Subject	press report
Whom/what?	Object	accident report
Where?	Locative	field mouse
When?	Time	night attack
Whose?	Possessive	family estate
What is it part of?	Whole-Part	duck foot
What are its parts?	Part-Whole	daisy chain
What kind of?	Equative	flounder fish
How?	Instrument	paraffin cooker
What for?	Purpose	bird sanctuary
Made of what?	Material	alligator shoe
What does it cause?	Causes	disease germ
What causes it?	Caused-by	drug death

82

Symbolic / Conceptual-based Systems (7)

Algorithm for applying rules:

- Rules check the semantic attributes to be satisfied;
- Groups of rules:
  - Modifier-based (e.g., subject-of, object-of, location-of, etc.)
  - Head-based (e.g., located-at, part-of, has,-part, hypernym, etc.)
  - Deverbal-based (e.g., has-subject, by-means-of, etc.);

83

Symbolic / Conceptual-based Systems (8)

Evaluation:

- Semantic information automatically extracted from LDOCE: 94,000 attribute clusters extracted, from nearly 75,000 single noun and verb definitions. Accuracy of 78%, with an error margin of +/- 5%.

- Training corpus:
  - 100 NSs from the examples in the previous literature (to make sure all noun classes are handled) (79%).
- Test corpus:
  - 97 NSs from the tagged version of the Brown corpus.
  - Accuracy: 52%.

84

Symbolic / Conceptual-based Systems (9)

Symbolic systems:

(Barker 1998):

- Semi-automatic, domain-dependent system
- Describes NCs as triplets of information (NMR):
  <modifier; head; marker>
- Relations initially assigned by hand; new ones assigned based on their similarity to previously classified NCs;
- Defines 50 semantic relation classes (uses 10);
- Deals with compositional noun compounds (meaning derived from the meaning of its elements);

85

Symbolic / Conceptual-based Systems (10)

- User interaction:
  - Initially there is not list of triples to match with the triple at hand. So, the user supplies the correct NMR when the system cannot determine it automatically.
  - User needs to be familiar with the NMR definitions (use paraphrases).

86

Symbolic / Conceptual-based Systems (11)

Assignming NMRs:

- Modifier-head
  (Marker=NIL; e.g., monitor cable);
- Postmodifying preposition
  (Marker = prep; e.g., pile of garbage);
- Appositives
  (Marker = appos; e.g., the dog, my best friend);

Distance between triples:

Dist.	Crt. Tripple	Prev. tripple	Example
0	(M, H, Mk)	(M, H, Mk)	"wall beside a garden" "wall beside a garden"
1	(M, H, <prep>	(M, H, nil)	"all beside a garden" "garden wall"
2	(M, H, <prep>)	(_, _, <prep>)	"pile of garbage" " house of bricks"

Symbolic / Conceptual-based Systems (12)

Evaluation:

- In the context of a large knowledge acquisition system.
- Criteria:
  - The analyzer's ability to learn to make better suggestions to the user as more NPs are analyzed.
  - Its coverage.
  - The burden the system places on the user.

Results:

- 886 modifier-noun pairs were assigned an NMR.
  - 608 (69%) were assigned correctly by the system.
    - For 97.5% the system offered a single suggestion.
  - After 100 assignments the system was able to make the majority of assignments automatically.

Symbolic / Conceptual-based Systems (13)

Symbolic systems:

(Hearst 1992, 1998):

"Automatic Acquisition of Hyponyms from Large Text Corpora"

- procedure for the automatic acquisition of the hypernymy lexical relation from unrestricted text;
- Identify a set of accurate lexico-syntactic patterns expressing hypernymy that:
  - Occur with high frequency in text;
  - Almost always represent the relation considered;
  - Can be recognized with little or no precoded knowledge;
- Suggests the same algorithm can apply to other semantic relations;

89

Symbolic / Conceptual-based Systems (14)

Procedure:

- Pick a semantic relation R;
- Get a list of terms between which R holds;
- Search automatically a corpus after the pairs of terms;
- Find what is common in these environments and hypothesize that common patterns would yield to the relation of interest;
- Use the patterns thus discovered to extract new instances of the relation considered;

90

Symbolic / Conceptual-based Systems (15)

Lexico-Syntactic Pattern	Example
NP0 such as NP1, NP2, .. NPn	"companies, such as IBM"
such NP0 as NP1 .. NPn	"songs by such singer as Bon Jovi.."
NP1…NPn or other NP0	"bruises, wounds, broken bones, or other injuries"
NP1…NPn and other NP0	"temples, treasuries, and other buildings"
NP1 including NP2	"common law countries, including Canada"
NP1, especially NP2	"European countries, especially France"

91

Symbolic / Conceptual-based Systems (16)

Symbolic systems:

(Girju, Badulescu, Moldovan 2003, 2006):

"Automatic Discovery of Part-Whole Relations. "

Goal: uncover the general aspects of NP semantics:

- What influences the semantic interpretation of various NP constructions?
- Is there only one interpretation model that works best for all types of expressions?
- What parameters govern the models capable of such semantic interpretation?

92

Symbolic / Conceptual-based Systems (17)

"The car's mail messenger is busy at work in the mail car as the train moves along. Through the open side door of the car, moving scenery can be seen. The worker is alarmed when he hears an unusual sound. He peeks through the door's keyhole leading to the tender and locomotive cab and sees the two bandits trying to break through the express car door."

Part(X, Y);

Q&A:
What are the components of Y?
What is Y made of?

Symbolic / Conceptual-based Systems (18)

The semantics of Meronymy:

- Complex relation that "should be treated as a collection of relations, not as a single relation" (Iris et al. 1988).
- Classification of part-whole relations: (Winston, Chaffin and Herman 1987)
  - Component – Integral (wheel – car);
  - Member – Collection (soldier – army);
  - Portion – Mass (meter – kilometer);
  - Stuff – Object (alcohol – wine);
  - Feature – Activity (paying – shopping);
  - Place – Area (oasis – desert);

Symbolic / Conceptual-based Systems (19)

Lexico-syntactic patterns expressing Meronymy:

Variety of meronymic expressions:

"The cloud was *made of* dust."
"Iceland is a *member of* NATO."
"The horn is *part of* the car."
(* "He is part of the game.")
"girl*'s* mouth",
"eyes *of* the baby",
"oxygen-rich water";

95

Symbolic / Conceptual-based Systems (20)

Previous work:

- Not much work done in automatic detection of Meronymy;
- (Hearst 1998):
  - Method for automatic acquisition of hypernymy (IS-A) relations based on a set of (mostly) unambiguous lexico-syntactic patterns;
- (Berland & Charniak 1999):
  - Statistical method on a very large corpus to find part-whole relations;
  - Input: list of wholes;
  - Output: ordered list of possible parts;
  - Accuracy: 55% (first 50 parts);

96

Symbolic / Conceptual-based Systems (21)

The approach

Supervised, knowledge intensive learning method;

- focus only on compositional compounds;

Phases:

1. Extraction of lexico-syntactic patterns expressing meronymy
2. Learning semantic constraints to identify part-whole relations

97

Symbolic / Conceptual-based Systems (22)

WordNet

- most widely-used lexical database for English
- free!
- G. Miller at Princeton (www.cogsci.princeton.edu/~wn)
- used in many applications of NLP
- includes entries for open-class words only (nouns, verbs, adjectives & adverbs) organized into hierarchies:
  - Eg: initially 9 hierarchies for nouns:

    abstraction, act, entity, event, group, phenomenon, possession, psychological feature, state;

  In newer versions all these hierarchies are linked under entity

98

Symbolic / Conceptual-based Systems (23)

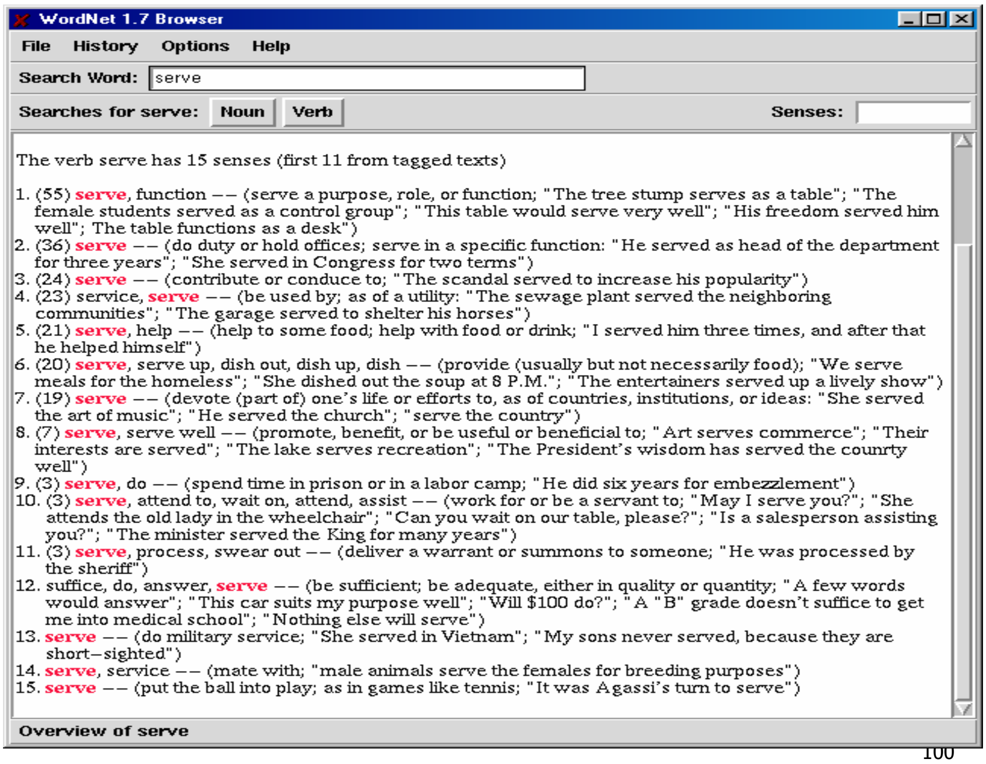
In WordNet 3.0:

- word forms organized according to their meanings (senses)

Category	Nouns	Synsets	Senses	polysemy
Nouns	117,097	81,426	145,104	1.23
Verb	11,488	13,650	24,890	2.16
Adjective	22,141	18,877	31,302	1.41
Adverb	4,601	3,644	5,720	1.24

- each entry has
  - a dictionary-style definition (gloss) of each sense
  - AND a set of domain-independent lexical relations among
    - WordNet's entries (words)
    - senses
    - sets of synonyms
- grouped into synsets (i.e. sets of synonyms)

99



100

Symbolic / Conceptual-based Systems (25)

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => life form, organism, being...
                  => entity, something
               => causal agent, cause, causal agency
                  => entity, something
Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument
   => instrument
      => device
         => instrumentality, instrumentation
            => artifact, artefact
               => object, physical object
                  => entity, something
```


Symbolic / Conceptual-based Systems (26)

- WordNet's Meronymy relations:
  - MEMBER-OF ("UK#1 – NATO#1")
  - STUFF-OF ("carbon#1 – coal#1")
  - PART-OF ("leg#3 – table#2")

  - Distributed over all nine noun hierarchies;

Symbolic / Conceptual-based Systems (27)

Phase I: An algorithm for finding lexico-syntactic patterns

(inspired from Hearst 1998)

- Step 1: Pick pairs of WordNet concepts Ci and Cj linked by a part-whole relation:
  - 100 pairs of part-whole concepts evenly distributed over all nine WordNet noun hierarchies and part-whole types;
- Step 2: Extract lexico-syntactic patterns linking each pair of concepts by searching a text collection:
  - SemCor 1.7 (10,000 sentences) and TREC-9 LA Times (10,000 sentences);
  - Lexico-syntactic patterns by manual inspection;

Symbolic / Conceptual-based Systems (28)

- Results:
  - 535 part-whole occurrences:
    - Phrase-level patterns: 493 (92.15%);
      - E.g.: eyes of the baby;
    - Sentence-level patterns: 42 (7.85%);
      - E.g.: the baby has blue eyes;
  - 42 distinct meronymic lexico-syntactic patterns:
    - 31 phrase-level patterns,
    - 11 sentence-level patterns;

Symbolic / Conceptual-based Systems (29)

Classification into four clusters based on the semantic similarity and frequency of occurrence of the part-whole patterns;

Cluster	Patterns	Freq.	Coverage [%]	Examples
C1	NPx of NPy NPx 's NPx NPy have NPx	282	52.71	"eyes of the baby" "girl's mouth" "the table has four legs"
C2	NPxy NPyx	86	16.07	"door knob" "apple pie"
C3	NPy PPx NPx PPy	133	24.86	"a bird without wings"
C4	Others	34	6.36	"In a car, the car body covers the engine."

105

Symbolic / Conceptual-based Systems (30)

Frequency of part-whole patterns:

- The frequent patterns are ambiguous. E.g.:
  - "Mary's toy" (Possession);
  - "Mary has a brother" (Kinship);

→ Need semantic constraints for pattern disambiguation;

106

Symbolic / Conceptual-based Systems (31)

Empirical observations on position of *part* and *whole*:

- Cluster 1:
  - of-genitives ("door of the car");
  - s-genitives ("car's door");
  - have-verb ("car has door");
- Cluster 2 ("car door", "ham sandwich");
- Cluster 3 ("seat in the car", "car with four doors");

107

Symbolic / Conceptual-based Systems (32)

Phase II: Learning semantic constraints (ISS – Iterative Semantic Specialization)

Approach
- Supervised, knowledge intensive learning method;
- Binary classification problem;
- C4.5 decision tree learning (Quinlan 1993);
  - Input:
    - positive and negative examples;
    - List of attributes with their possible values and classes;
  - Output:
    - Learned function represented by a decision tree or a set of if-then rules;

108

Symbolic / Conceptual-based Systems (33)

- Machine Learning – Definition

"The subfield of AI concerned with programs that learn from experience"
Russell / Norvig, AIMA

Symbolic / Conceptual-based Systems (34)

- Example:

Training Set

Outlook	Temperature	Humid.	Wind	Play?
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rain	High	Mild	FALSE	Yes

- Use this training set to learn how to classify new days:

Outlook	Temp.	Humid.	Wind	Play?
Sunny	Mild	High	FALSE	?
Sunny	Cool	Normal	FALSE	?
Rain	Mild	High	TRUE	?

Test Set

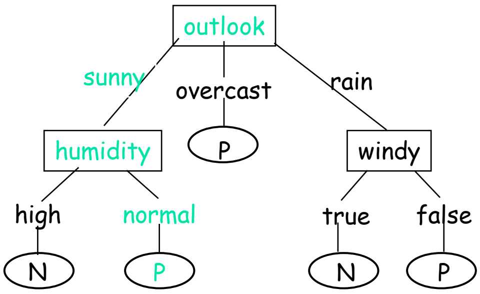
Input Data

Classification

- The input data is often easily obtained, whereas the classification is not.

Symbolic / Conceptual-based Systems (35)

- One rule is generated for each path in the tree from the root to a leaf
- Rules are generally simpler to understand than trees



IF outlook=sunny
AND humidity=normal
THEN play tennis

111

Symbolic / Conceptual-based Systems (36)

- Goal: Use training set + some learning method to produce a predictive model.
- Use this predictive model to classify new data.
- Sample applications:

112

Symbolic / Conceptual-based Systems (37)

Learning semantic constraints:

Preprocessing part-whole lexico-syntactic patterns:

- Preprocess NPs to identify the *part* and *whole* concepts cf. WordNet;
- For each NP keep only the largest word sequence from left to right in WordNet:
  - "brown carving knife" => "carving knife" (WordNet);
- Manually semantically disambiguate the training concepts:
  - carving_knife#1

113

Symbolic / Conceptual-based Systems (38)

Building the Training corpus:

- SemCor 1.7 and TREC-9 LA Times collections;
- Corpus A:
  - 19,000 sentences (SemCor 1.7);
  - 100,000 sentences (LA Times);
  - Syntactically parsed (Charniak 2000);
- Focus only on sentences matching the patterns considered and perform manual semantic disambiguation (exception SemCor);

114

Symbolic / Conceptual-based Systems (39)

WordNet part-whole relations searched on Altavista;
Query example: "door" AND "knob"

Cluster	Positive examples		Negative examples (corpus "A")
	From WordNet	From corpus "A"	
C1	27,636	325	18,611
C2	142	625	6,601
C3	111	295	2,751

115

Symbolic / Conceptual-based Systems (40)

Inter-Annotator Agreement:

- 2 graduate students in Computational Semantics;
- Instructions:
  - Input: sentences containing pair;
    WCH part-whole classification;
  - Output: word senses for pair nouns;
    position of *part* and *whole*;
    annotation of pair as part-whole/not;

3rd judge decided on the non-agreed word senses and relations;

116

Symbolic / Conceptual-based Systems (41)

Inter-Annotator Agreement

$$K = \frac{Pr(A) - Pr(E)}{1 - Pr(E)}$$

Type of example	Cluster	Kappa Agreement
Training	C1	0.878
	C2	0.826
	C3	0.811
Testing	C1	0.880
	C2	0.862
	C3	0.836

117

Symbolic / Conceptual-based Systems (42)

The ISS Learning Algorithm:

- *Input*: positive and negative meronymic examples of pairs of concepts;
- *Output*: semantic constraints on concepts;

- Step 1: Generalize the training examples:
  - Initial corpus:

    <part#sense; whole#sense; target>

    E.g.: <academician#1; academy#2; Yes>

118

Symbolic / Conceptual-based Systems (43)

- Intermediate corpus:

  <part#sense; class_part#sense;
  whole#sense; class_whole#sense;
  target>

<academician#1; academy#2; Yes>

<academician#1, **entity**#1; academy#2, **group**#1; Yes>

119

Symbolic / Conceptual-based Systems (44)

- Generalized training corpus (example types):
  - Positive examples:

    <X_hierarchy#sense; Y_hierarchy#sense; Yes>
  - Negative examples:

    <X_hierarchy#sense; Y_hierarchy#sense; No>

120

Symbolic / Conceptual-based Systems (45)

- Ambiguous examples:

<X_hierarchy#sense; Y_hierarchy#sense; Yes/No>

<apartment #1; woman #1; No>
<hand #1; woman #1; Yes>

<**entity** #1; **entity** #1; Yes/No>

Symbolic / Conceptual-based Systems (46)

- Step 2: Learning constraints for unambiguous examples:
  - Form new corpus only from unambiguous examples;
  - apply C4.5 using 10-fold cross validation;
  - Rank rules according to their frequency of occurrence and average accuracy obtained for each particular set;
  - Keep rules with freq. > 70% and average accuracy > 50%;

Symbolic / Conceptual-based Systems (47)

- Step 3: Specialize the ambiguous examples:
  - Based on IS-A information provided by WordNet
  - Initially: semantic class as the root of the noun hierarchies in WordNet;
  - Specialization: replace each semantic class with its corresponding hyponym;
- Repeat steps 2 and 3 until there are not more ambiguous examples;

Symbolic / Conceptual-based Systems (48)

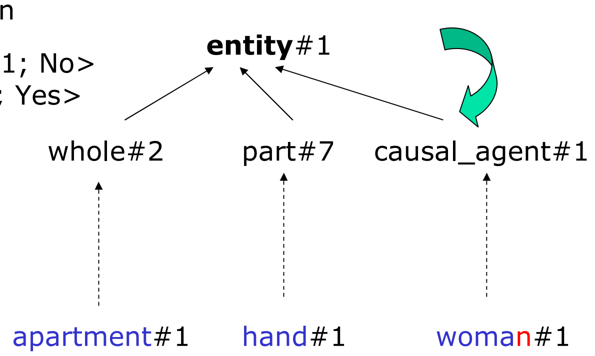
Specialization example1:

woman#1 's apartment#1;
hand#1 of a woman#1;

<entity#1; entity#1; Yes/No>

specialization

<whole#2; causal_agent#1; No>
<part#7; causal_agent#1; Yes>



Symbolic / Conceptual-based Systems (49)

Specialization example2:

<entity#1; entity#1; Yes/No>

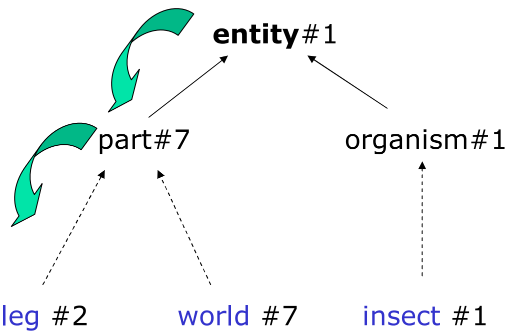
⇓ specialization

<part#7; organism#1; No>
<part#7; organism#1; Yes>

⇓ specialization

. . .

leg#2 of insect#1;
insect#1 's world#7;



125

Symbolic / Conceptual-based Systems (50)

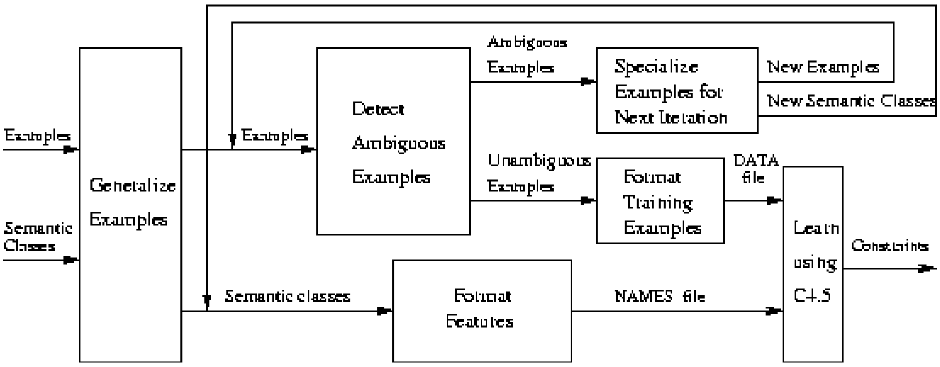
- Example of iterative specializations:

entity#1 – entity#1
    body_of_water#1 – location#1
    location#1 – body_of_water#1
    location#1 – object#1
        region#1 – artifact#1
        region#3 – artifact#1

126

Symbolic / Conceptual-based Systems (51)

System architecture



127

Symbolic / Conceptual-based Systems (52)

- A sample of semantic constraints

Part **Class**	***Whole*** **Class**	**Val.**	**Acc. [%]**	**Freq.**	**Example**
shape#2	artifact#1	1	69.84	9	*point#8 - knife#2 (knife point)*
process#1	person#1	0	76.70	8	*growth#2 – child#2 (growth of child)*
person#1	person#1	0	89.55	10	*child#1 – woman#1 (woman’s child)*
location#1	object#1	1	85.64	10	*base#16 – box#1 (base of a box)*

- 27 sets of constrains for Cluster 1;

128

Symbolic / Conceptual-based Systems (53)

Performance metrics

$$\text{Precision} = \frac{\text{No. of correctly retrieved relations}}{\text{No. of relations retrieved}}$$

$$\text{Recall} = \frac{\text{No. of correctly retrieved relations}}{\text{No. of correct relations}}$$

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

129

Symbolic / Conceptual-based Systems (54)

Test corpus

- other 10,000 sentences of TREC-9 LA Times news articles:
  - 4,106 (C1 patterns)
  - 3,442 (C2 patterns)
  - 2,577 (C3 patterns)

- Syntactically parsed and disambiguated with a WSD system (Mihalcea & Moldovan 2001);
- NE Recognizer (96% F-measure on MUC-6 data);

130

Symbolic / Conceptual-based Systems (55)

Experiments with Cluster 2:

- Use set of constraints obtained for:
  - C1 ("X of Y", "Y's X", "Y has X")
  - C1 + C2 (.., "XY", "YX")
  - C2 ("XY", "YX")

Results	C1 [%]	C1 + C2 [%]	C2 [%]
Precision [%]	48.43	52.98	**79.02**
Recall for cluster [%]	58.08	73.46	**75.33**
F-measure	52.82	61.56	**77.13**

131

Symbolic / Conceptual-based Systems (56)

Experiments with Cluster 3

- Use set of constraints obtained for:
  - C1 ("X of Y", "Y's X", "Y has X")
  - C2 ("XY", "YX")
  - C1 + C3 (.., "X prep Y", "Y prep X")
  - C2 + C3 (.., "X prep Y", "Y prep X")
  - C3 ("X prep Y", "Y prep X")
  - C1 + C2 + C3

Results	C1 [%]	C2 [%]	C1 + C3 [%]	C2 + C3 [%]	C3 [%]	C1 + C2 + C3 [%]
Precision [%]	48.43	61.54	4.81	36.36	**82.56**	40.74
Recall for cluster [%]	58.08	54.55	8.26	36.36	**62.83**	15.06
F-measure	52.82	57.84	6.18	36.36	**71.36**	22.78

132

Symbolic / Conceptual-based Systems (57)

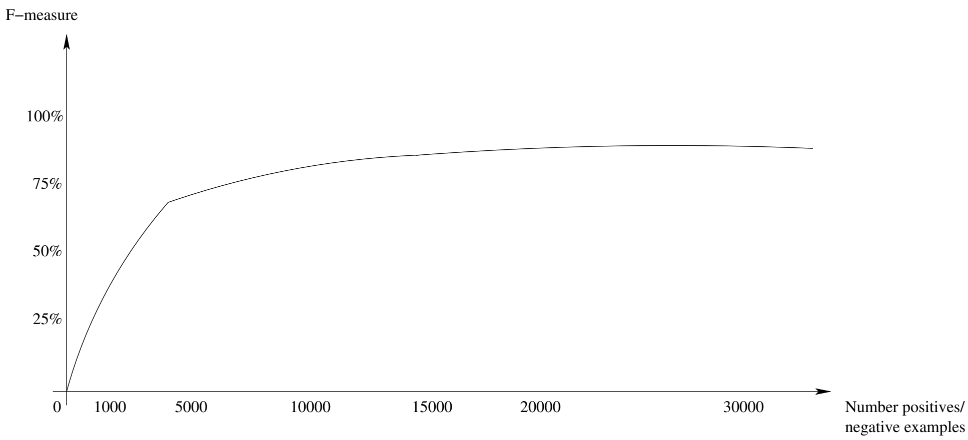
Overall results

Results	Baseline1 (No WSD) [%]	Baseline2 (One learning) [%]	Baseline3 (No generaliz.) [%]	Baseline4 (autom. WSD) [%]	Part-Whole system [%]
Precision	7.72	7.73	15.71	53.57	**82.87**
Recall for patts.	24	43	2.95	27.87	**79.09**
Coverage	10.81	19.37	2.71	25.86	**72.66**
F-measure	3.56	6.02	4.97	36.67	**82.05**

133

Symbolic / Conceptual-based Systems (58)

Learning curve



134

Symbolic / Conceptual-based Systems (59)

Conclusions

- Advantages:
  - Training data – not noisy;
  - WordNet IS-A hierarchy + WSD (their importance);
  - Good model:
    - 75% F-measure with only 16.8% training data;
    - Independent of patterns used;
  - Linguistic evidence: cluster of patterns have different semantic behavior;

- Disadvantages:
  - Heavily dependent on WordNet and WSD;
  - Requires a very large annotated corpus for training;

135

Symbolic / Conceptual-based Systems (60)

System improvements

Semantic Scattering (SS)

(Moldovan at al. 2004) (Badulescu & Moldovan 2005) (Girju at al. 2005)

Semantic Scattering2 (SS2) (Beamer, Girju, Rozovskaya, 2008)

- Extended the system to cover other semantic relations (22 semantic relations list)
- Changed the learning model from ISS to SS and then to SS2

136

Symbolic / Conceptual-based Systems (61)

Semantic Scattering

The Data and Inter-Annotator Agreement

Source	LA Times/TREC 9
Size	20,000 sentences
Of-genitives	2,249
S-genitives	1,006
Inter-annotator agreement	82%
Training/Development/ Testing	80/10/10

137

Semantic Relation	Of-Genitives	S-Genitives
POSSESSION	36	220
KINSHIP	25	61
PROPERTY	109	75
AGENT	11	123
TEMPORAL	5	109
DEPICTION	30	7
PART-WHOLE	328	114
IS-A	0	0
CAUSE	10	3
MAKE/PRODUCE	11	62
INSTRUMENT	0	0
LOCATION/SPACE	32	46
PURPOSE	0	0
SOURCE	56	33
TOPIC	70	5
MANNER	0	0
MEANS	0	0

Semantic Relation	Of-Genitives	S-Genitives
ACCOMPANIMENT	10	4
EXPERIENCER	1	2
RECIPIENT	49	41
ASSOCIATED WITH	5	2
MEASURE	115	1
THEME	120	50
RESULT	8	2
OTHER	107	49

138

Symbolic / Conceptual-based Systems (63)

Features:

- Semantic class of head noun: f_j^h
  - *child's mother* [KINSHIP]
  - *child's toy* [POSSESSION]
- Semantic class of modifier noun: f_i^m
  - *Mary's apartment* [POSSESSION]
  - *apartments of New York* [LOCATION]

- Feature pair: $<f_i^m, f_j^h> = f_{ij}$

- Form tuples: $<f_{ij}, r>$

139

Symbolic / Conceptual-based Systems (64)

Observations:

- f_i^m and f_j^h can be regarded as nodes on some paths that link the senses of the most specific noun concepts with the top of the noun hierarchies.

- The closer the pair of noun senses f_{ij} is to the bottom of noun hierarchies the fewer the semantic relations associated with it; the more general f_{ij} is the more semantic relations.

$$P(r \mid f_{ij}) = \frac{n(r, f_{ij})}{n(f_{ij})}$$

140

Symbolic / Conceptual-based Systems (65)

The model

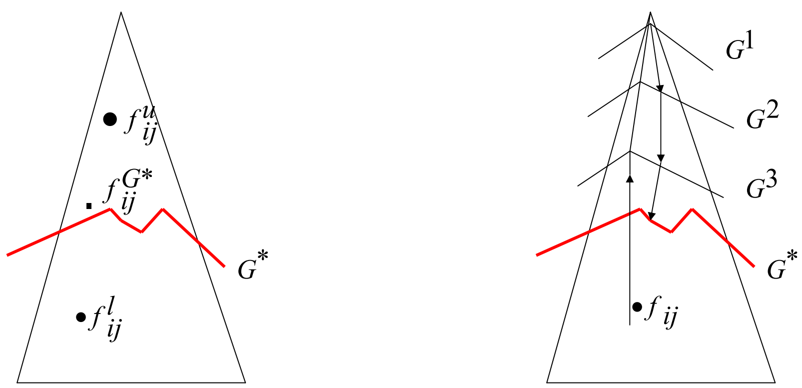
- Case 1: f_{ij} is specific enough such that there is only one semantic relation r for which $P(r \mid f_{ij}) = 1$ and the rest 0.
- Case 2: There are more than two semantic relations for which $P(r \mid f_{ij})$ is different from 0.

$$\hat{r} = \arg\max P(r \mid f_{ij})$$

141

Symbolic / Conceptual-based Systems (66)

Conceptual view of the noun hierarchy separated by boundary G*



142

Symbolic / Conceptual-based Systems (67)

Boundary Detection Algorithm:

- Step 1: Create an initial boundary G^1
- Step 2: Specialize the Boundary G^1
  - 2.1 Construct a lower boundary
  - 2.2 Test the new boundary
  - 2.3 Repeat steps 2 and 3 till there is no more performance improvement.

143

Symbolic / Conceptual-based Systems (68)

- Sample row from the conditional probability table

R	1	2	3	6	7	11	13	15	16	19	21	24	25	Others
P*	0.048	0.120	0.006	0.032	0.430	0.016	0.035	0.285	0.012	0.004	0.010	0.001	0.001	0

P* = P(r | entity – entity)

144

Symbolic / Conceptual-based Systems (69)

Statistics for the Semantic Class Features by Level of Specialization

	Of-genitives			S-genitives		
	G^1	G^2	G^3	G^1	G^2	G^3
No. modifier features	9	31	74	9	37	91
No. head features	9	34	66	9	24	36
No. feature pairs	63 out of 81	216 out of 1054	314 out of 4884	41 out of 81	157 out of 888	247 out of 3276
No. features with one relation	26	153	281	14	99	200
Avg. number non zero rels per line	3	1.46	1.14	3.59	1.78	1.36

145

Symbolic / Conceptual-based Systems (70)

Semantic Relation Classification Algorithm

- Step 1. Process the sentence
- Step 2. Identify the head and modifier noun concepts
- Step 3. Identify the feature pair
- Step 4. Find the semantic relation

146

Symbolic / Conceptual-based Systems (71)

	Of-genitives			S-genitives		
	Baseline1	**Baseline2**	**Results**	**Baseline1**	**Baseline2**	**Results**
Number of correctly retrieved relations	49	59	81	15	27	71
Number of relations retrieved	73	75	99	63	66	85
Number of correct relations	104	104	104	96	96	96
Precision	67.12%	76.62%	**81.82%**	23.81%	40.91%	**83.53%**
Recall	47.12%	56.73%	**77.88%**	15.63%	28.13%	**73.96%**
F-measure	55.37%	65.92%	**79.80%**	18.87%	33.34%	**78.45%**

147

Symbolic / Conceptual-based Systems (72)

Comparison with other models

Accuracy	Of-genitives	S-genitives
Semantic Scattering	**79.85%**	**78.75%**
Decision Trees (C5.0)	40.60%	47.0%
Naïve Bayes (JBNC)	42.31%	43.7%
SVM (LibSVM)	31.45%	23.51%

148

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Applications of semantic relations
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

149

Context-based Systems (1)

SemEval 2007

- 4th International Workshop on Semantic Evaluations
  - year-long activity for task organizers
  - start: May 2006 – call for task proposals
  - finish: June 2007 – workshop at ACL 2007
  - 27 task proposals, 19 accepted, 16 completed
- Task 04
  - Classification of Semantic Relations between Nominals
  - organizers: Roxana Girju, Marti Hearst, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, Deniz Yuret
  - 14 participating teams

150

Context-based Systems (2)

	Task	Papers	Name of Task
1	**17**	**22**	**English Lexical Sample, English SRL and English All-Words Tasks**
2	**13**	**15**	**Web People Search**
3	**04**	**14**	**Classification of Semantic Relations between Nominals**
4	**07**	**11**	**Coarse-grained English all-words**
5	**10**	**8**	**English Lexical Substitution for SemEval-2007**
6	**05**	**6**	**Multilingual Chinese-English Lexical Sample**
7	**15**	**6**	**TempEval: Evaluating Time-Event Temporal Relation Identification**
8	**02**	**5**	**Evaluating Word Sense Induction and Discrimination Systems**
9	**14**	**5**	**Affective Text**
10	**08**	**5**	**Metonymy Resolution at Semeval-2007**
11	**19**	**4**	**Frame Semantic Structure Extraction**
12	**01**	**3**	**Evaluating WSD on Cross Language Information Retrieval**
13	**06**	**3**	**Word-Sense Disambiguation of Prepositions**
14	**11**	**2**	**English Lexical Sample via English-Chinese Parallel Text**
15	**09**	**2**	**Multilevel Semantic Annotation of Catalan and Spanish**
16	**18**	**1**	**Arabic Semantic Labeling**

Context-based Systems (3)

Application: Relational Search

Relation	Example of a relational search query
Cause-Effect	**"list all *X* such that *X*** causes **cancer"**
Part-Whole	**"list all *X* such that *X*** is part of **an automobile engine"**
Material-Object	**"list all *X* such that *X*** is material for **making a ship's hull"**
Hypernym-Hyponym	**"list all *X* such that *X*** is a type of **transportation"**
Origin-Entity	**"list all *X* such that *X*** is produced from **cork trees"**

Context-based Systems (4)

SemEval Task 4:

- Focus on:
  - nominals
  - a number (7) of interesting relations;
- Relations between nominals:
  - As noun – modifier relations / noun compounds:
    - summer holidays, flu virus,
  - As lexical relations:
    - Cat – feline, day - time
- Between nominals in a sentence:

> "Among the contents of the **vessel** were a set of carpenter's **tools**, …"

153

Context-based Systems (5)

Relations:

- **Cause-Effect**: laugh wrinkles
- **Instrument-Agency**: laser printer
- **Product-Producer**: honey bee
- **Origin-Entity**: message from outer-space
- **Theme-Tool**: news conference
- **Part-Whole**: car door
- **Content-Container**: the air in the jar

154

Context-based Systems (6)

Relation specifications

Content-Container

- **Definition:** Content-Container(X, Y) is true … iff:
  … X is or was (usually temporarily) stored or carried inside Y**;**
- **Definition – restrictions:** The container is clearly delineated in space; …
- **Definition – notes:** The relation applies to concrete and abstract content; …
- **Positive examples:** The *apples* are in the *basket*.
- **Near-miss negative examples:** The *apples* were stacked in a *pyramid*.

155

Context-based Systems (7)

Datasets creation and annotation

- Extract sentences by querying using patterns (Annotator 1);
  Content-Container: "the * in the *", "* contains *", …

156

Context-based Systems (8)

Datasets creation and annotation

- Extract sentences by querying using patterns (Annotator 1);
- Tag entities (Annotator 1);

After the cashier put the <e1>**cash**</e1> in a <e2>**bag**</e2>, the robber saw a bottle of scotch that he wanted behind the counter on the shelf.

Query = “the * in a *”

157

Context-based Systems (9)

Datasets creation and annotation

- Extract sentences by querying using patterns (Annotator 1);
- Tag entities (Annotator 1);
- Annotate entities with WN senses (Annotator 2, Annotator 3);

After the cashier put the <e1>**cash**</e1> in a <e2>**bag**</e2>, the robber saw a bottle of scotch that he wanted behind the counter on the shelf.

Query = “the * in a *”
WordNet(e1) = "cash%1:21:00::"
WordNet(e2) = "bag%1:06:00::"

158

Context-based Systems (10)

Datasets creation and annotation

- Extract sentences by querying using patterns (Annotator 1);
- Tag entities (Annotator 1);
- Annotate entities with WN senses (Annotator 2, Annotator 3);
- Annotate relation (True/False) (Annotator 2, Annotator 3);

After the cashier put the <e1>**cash**</e1> in a <e2>**bag**</e2>, the robber saw a bottle of scotch that he wanted behind the counter on the shelf.

Query = "the * in a *"
WordNet(e1) = "cash%1:21:00::"
WordNet(e2) = "bag%1:06:00::"
Content-Container(e1,e2) = "true"

159

Context-based Systems (11)

Datasets

- 140 training and >70 test examples for each relation;
- Balanced positive and negative examples.

- Annotation summary:
  - WN sense inter-annotator agreement: 71.9%
  - Initial relation inter-annotator agreement: 70.3%
  - Disagreements discussed and consensus reached (or example thrown out).
  - Definitions modified to reflect discussed examples.

160

Context-based Systems (12)

Competition:

- 14 teams
- 15 systems

- System set-up variations:
  - WN senses (yes/no)
  - Query (yes/no)
  - 4 levels of training data (35,70,105,140)

161

Context-based Systems (13)

Learning algorithms

	Teams	Supervised Learning Algorithms
1	**UVAVU**	**Logistic Regression, *k* Nearest Neighbours, Decision Trees**
2	**CMU-AT**	**SVM-light (Support Vector Machine) with tree-kernel package**
3	**ILK**	**Naïve Bayes, Bayes Nets, Decision Stumps, LWL, IBk, IB1, J48**
4	**FBK-IRST**	**five kernel functions (words, contexts, parse trees, senses, synonyms)**
5	**LCC-SRN**	**Decision Trees (C5), SVM (libSVM), Naïve Bayes, MaxEnt, SS, ISS**
6	**UMELB**	**Semantic Scattering with Maximum Likelihood, 1NN with WordNet Sim**
7	**UC3M**	**Sequential Minimal Optimization (a type of Support Vector Machine)**
8	**UCD-S1**	**Sequential Minimal Optimization (a type of Support Vector Machine)**
9	**UCD-FC**	**Naïve Bayes**
10	**UCD-PN**	**Sequential Minimal Optimization (a type of Support Vector Machine)**
11	**UIUC**	**libSVM (open source Support Vector Machine) with RBF kernel**
12	**UTD-HLT-CG**	**Decision Trees, Decision Rules, Logistic Regression, Lazy Classifiers**
13	**UTH**	**TinySVM (Support Vector Machine) with linear kernel**
14	**UCB**	**Single Nearest Neighbour (1NN) with Dice coefficient for nearness**

162

Context-based Systems (14)

Summary of system approaches:

- Systems that used only info about the two entity nouns:
  - MELB-KB
  - UCD-PN (WN + web; no sentence features)
  - UCD-FN

- Systems that used info about the two nouns:
  - generalizations on IS-A hierarchy:
    - top k levels (CMU-AT, **FBK-IRST**, UCD-S1, UCD-PN, SPAIN) + lexicographer files (**FBK-IRST**, ILK, SPAIN, UTH)
  - similarity with training: usually WN similarity (UTH)

163

Context-based Systems (15)

Summary of system approaches:

- Systems using the web only:
  - UCD

- Systems that use additional corpora:
  - UIUC
  - UCD-FC
  - UCD-S1
  - UTH

164

Context-based Systems (16)

UIUC's Approach

- extend over our previous work:

  syntax ⟷ semantics

  - noun compounds and other noun phrases
    - E.g.: tea cup vs. cup of tea vs. *tea's cup
  - "noun verb noun" constructions
    - E.g.: wars lead to destruction vs. the path leads to the house

- employ a knowledge-rich supervised learning model

165

Context-based Systems (17)

Feature vector

- Feature set #1 (Core features)
  - Consider only the morpho-lexico-semantic information of the target nouns; (Girju et al. 2003, 2004, 2005, 2006)

- Feature set #2 (Context features)
  - Consider the sentence context in order to identify features at different linguistic levels;

- Feature set #3 (Special features)
  - Consider features that help identify specific information about some new semantic relations;

166

Context-based Systems (18)

Feature set #1: Core features

- Argument position (F1):
  - Position of semantic arguments in the relation;
  - E.g.: silk/e1 dress/e2 STUFF-OBJECT / PART-WHOLE(e1, e2)
    dress/e1 of silk/e2 STUFF-OBJECT / PART-WHOLE(e2, e1)
- Semantic specialization (F2):
  - Binary feature
  - Prediction of a semantic specialization learning model (series of semantic specialization procedures on WordNet);
  - E.g.: wars – destruction and path – house map to entity – entity; through specialization:
    abstraction – abstraction and physical entity – physical entity;
    (Girju et al. 2003, 2006)

167

Context-based Systems (19)

- Nominalization (F3, F4)
  - If target noun (e1 or e2) is a nominalization;
  - Values: *agential nouns*, *other nominalizations*, *neither;*
  - Resources: WordNet, NomLex-Plus (Meyers et al. 2004);
  - Rationale: to filter out some examples (e.g., car owner THEME)
- Spatio-Temporal features (F5, F6):
  - Resources: WordNet noun classes; special classes (Herskovits, 87), (Linstromberg, 97), (Tyler & Evans, 03), (Borcev, Rau, Partee 2004)
  - Rationale: to filter out some near-miss examples (e.g., activation *by* summer TMP or mouse *in* the field LOC)

168

Context-based Systems (20)

Feature set #2: Context features

- Grammatical role (F7, F8)
  - Values: *subject, direct object, neither*
  - Rationale: filter out some instances with poor context (e.g., noun compounds) and identify near-miss examples (e.g., THEME-TOOL)
  - E.g.: Mary gave the <e1>*girl*</e1> <e2>*books*</e2> with nice pictures.

- PP attachment (F9)
  - Applies to NP PP
  - Rationale: identify negative instances where PP attaches to any other word but the target noun (cf. Charniak 2000).
  - E.g.: John eats <e1>*pizza*</e1> with <e2>*a fork*</e2>.

169

Context-based Systems (21)

- Semantic role (F10, F11)
  - Binary feature which identifies the semantic role of a phrase in a verb – argument structure (the phrase contains either *e1* or *e2*)
  - Values: *Time, Location, Manner*
  - Rationale: filter out near-miss examples (e.g., Instrument – Agency relation);
  - Resources: ASSERT (Univ. of Colorado at Boulder)

170

Context-based Systems (22)

- Inter-noun context sequence (F12, F13, F14)
  - Encodes the sequence of lexical and POS information in between the two target nouns;
  - F14
    - is a weight feature on the values of F12 and F13
    - indicates how similar a new sequence is to the already observed inter-noun context associated with the relation;
    - If direct match => value = 1;
    - If POS of new substring matches that of an already seen substring => value = 0.5;
    - If sequence overlaps entirely/partially with patterns encoding other semantic relations in the same contingency set => 0.25 or 0.125
    - The value of the feature is the sum of the weights thus obtained;
  - Rationale: the greater the weight, the more representative is the context sequence for that relation

171

Context-based Systems (23)

Feature set #3: Special features

- Psychological feature (F15, F16) (for Theme-Tool):
  - Indicates if target noun belongs to a list of special concepts
  - Obtained from the restrictions listed in the definition Theme-Tool
  - E.g.: need for money (*need* is a psychological feature)
  - Resources: WordNet subhierarchies (motivation, cognition); preconditions (foundation, requirement);

- Instrument semantic role (F17) (for Instrument-Agency):
  - Binary feature
  - Indicates if argument identified as Instrument in the relation belongs to an instrument phrase (cf. ASSERT);

- Syntactic attachment (F18)
  - Indicates if the argument identified as Instrument in the relation attaches to a verb/noun in the syntactically parsed sentence

172

Context-based Systems (24)

Learning model and experimental setting:

- libSVM package (with RBF kernel)
- A binary classifier for each of the 7 relations

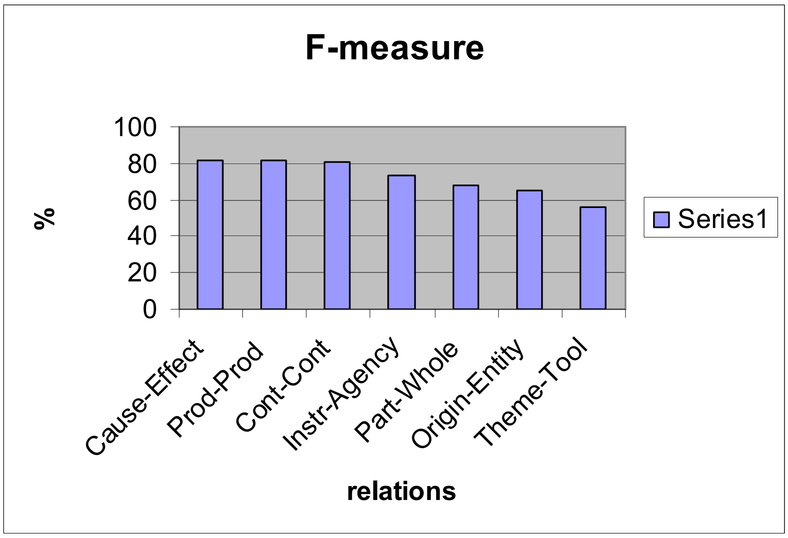
- Expansion of training data:
  - 1,320 Cause-Effect, 721 Product-Producer (news articles) (Girju 2003)
  - 1,003 Part-Whole, 167 Origin-Entity, 112 Product-Producer, 91 Theme-Tool (Wall Street Journal) (Moldovan et al. 2004), (Girju et al. 2005)
  - 552 Product-Producer (eXtended WordNet)
  - Theme-Tool and Content-Container (special lists: thesauri, selectional restrictions on Containers & Locations)

Context-based Systems (25)

Experimental results

Relation	P	R	F	Acc	Total	Base-F	Base-Acc
Cause-Effect	69.5	**100**	**82**	77.5	80	67.8	51.2
Instrument-Agency	**68.2**	78.9	73.2	71.8	78	65.5	51.3
Product-Producer	84.5	79	81.7	76.3	93	80	66.7
Origin-Entity	86.4	52.8	65.5	75.3	81	61.5	55.6
Theme-Tool	85.7	41.4	**55.8**	73.2	71	58	59.2
Part-Whole	70.8	65.4	68	77.8	72	53.1	63.9
Content-Container	**93.1**	71.1	80.6	82.4	74	67.9	51.4
Average	79.7	71.1	80.6	82.4	74	67.9	51.4

Context-based Systems (26)



175

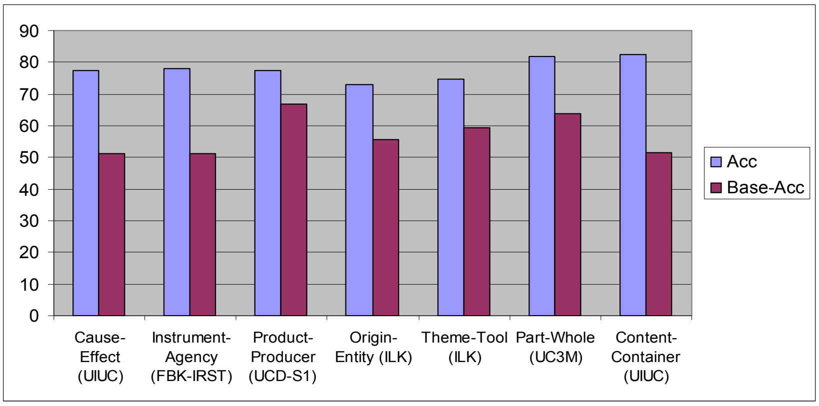
Context-based Systems (27)

- The correspondence syntax – semantics is definitely important,
  - however more information is needed for noun – noun relations beyond NP level;
- More data helps
  - but to what extent? (does it differ per semantic relation, genre, etc.?)
- Ad-hoc special resources help
  - However, much deeper research is needed to understand these relations and, thus to be able to build such resources (semi)automatically
  - E.g.: Content-Container, etc.

176

Context-based Systems (28)

SemEval 1007 Task 4: Best results per relation

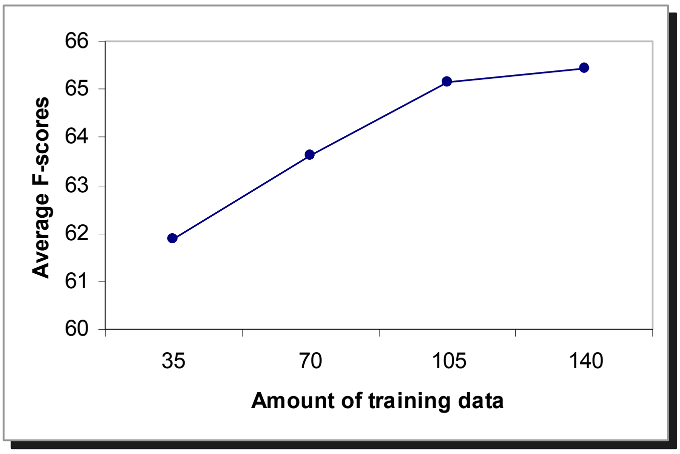


14 teams, 15 systems, Almost all winners B4: WN=Y, Qry=N, all data used

177

Context-based Systems (29)

Analysis: Does more data help?



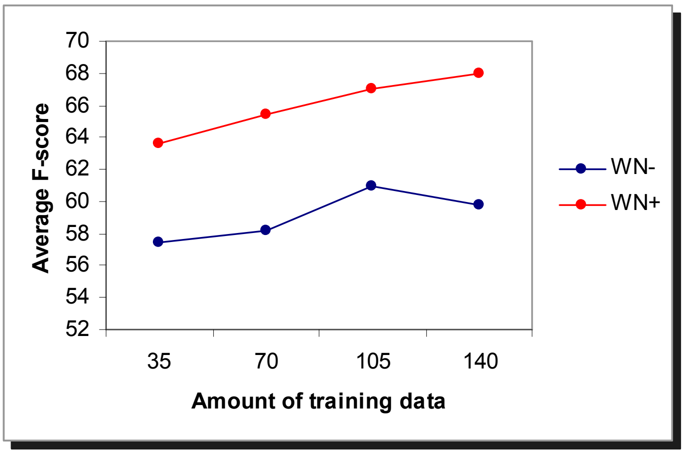
Average for 17 system/set-up combinations

178

Context-based Systems (30)

Analysis:

Does WordNet help?



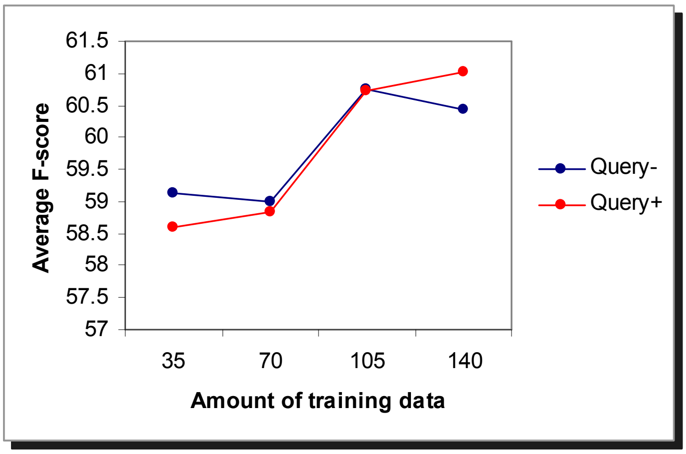
Comparison based on 3 systems that used both :
set-up A (Query = NO, WN = NO)
set-up B (Query = NO, WN = YES)

179

Context-based Systems (31)

Analysis:

Do queries help?

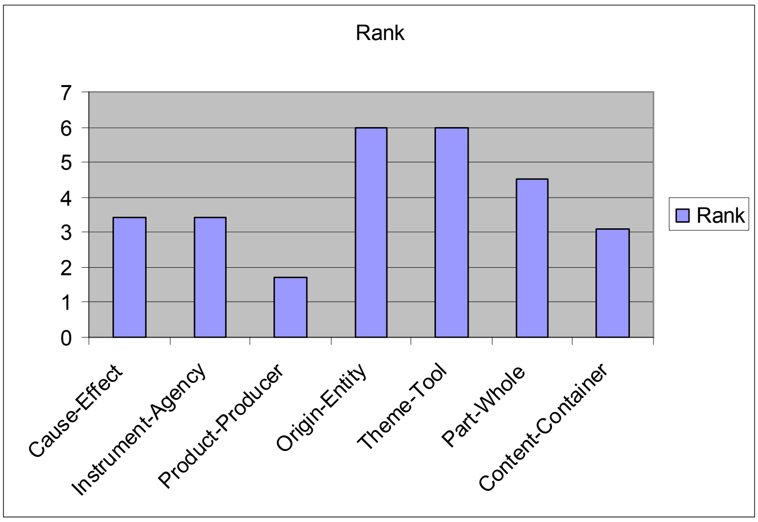


Comparison based on 3 systems that used both :
set-up A (Query = NO, WN = NO)
set-up C (Query = YES, WN = NO)

180

Context-based Systems (32)

Some relations are harder to classify



181

Context-based Systems (33)

Conclusions

- New task for classification of semantic relations
- Did not adopt a fixed inventory of semantic relations
- Focus on similarity of relations between different word groups
- More training data and WN senses help, queries do not
- Results significantly above baseline

182

Outline

183

Noun Compound Bracketing (1)

Problem definition

(a) [[liver cell] antibody] (left bracketing)

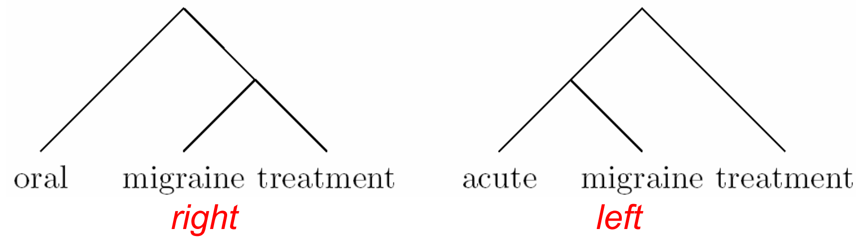
(b) [liver [cell line]] (right bracketing)

- In (a), the *antibody* targets the *liver cell*.
- In (b), the *cell line* is derived from the *liver*.

184

Noun Compound Bracketing (2)

Dependency vs. Adjacency



(oral (migraine treatment)) ((acute migraine) treatment)

oral migraine treatment

acute migraine treatment

dependency model **adjacency model**

185

Noun Compound Bracketing (3)

Previous work

- Marcus(1980), Pustejosky&al.(1993), Resnik(1993)
  - adjacency model: $Pr(w_1|w_2)$ vs. $Pr(w_2|w_3)$
- Lauer (1995)
  - dependency model: $Pr(w_1|w_2)$ vs. $Pr(w_1|w_3)$

Pr that w_1 precedes w_2

oral migraine treatment

acute migraine treatment

dependency model adjacency model

186

Noun Compound Bracketing (4)

- Keller & Lapata (2004):
  - use the Web
  - unigrams and bigrams
- Nakov & Hearst (2005):
  - use the Web
  - *n*-grams
  - paraphrases
  - surface features
- Girju at al (2005)
  - Supervised model
  - From meaning to bracketing: how does the semantics NCs help in their bracketing

187

Noun Compound Bracketing (5)

Lapata & Keller 2004

- compared:
  - Web-based n-grams (unsupervised)
  - BNC (smaller corpus)-based n-grams
- the Web-based n-grams were:
  - The same as or better than BNC-based n-grams
- Thus, they propose using the Web as a baseline against which most algorithms should be compared.

188

Noun Compound Bracketing (6)

Algorithm: Computing Web n-grams

- Find # of hits for each term via Altavista
  - This gives doc frequency, not term frequency
  - Smooth 0 counts to 0.5
  - All terms lower case

189

Noun Compound Bracketing (7)

Algorithm: Computing Web n-grams

- Three different types:
  - NEAR queries
    - A NEAR b: within a 10-word window
  - Inflected queries
    - Expand each term to all morphological variations
      - "history change"
      - "histories change"
      - "history changes" …

190

Noun Compound Bracketing (8)

- Which way do the nouns group?
  - Current best model (Lauer'95) uses a thesaurus and taxonomy in an unsupervised manner
- Method:
  - Compare the probability of left-branching to probability of right-branching (as done in Lauer'95)
  - However, use the Web counts to estimate the probabilities
    - Inflected queries and NEAR operator
- Results:
  - Far better than baseline; no significant difference from best model

Model	Alta	BNC
Baseline	63.93	63.93
$f(n_1,n_2) : f(n_2,n_3)$	77.86	66.39
$f(n_1,n_2) : f(n_1,n_3)$	78.68#∗	65.57
$f(n_1,n_2)/f(n_1) : f(n_2,n_3)/f(n_2)$	68.85	65.57
$f(n_1,n_2)/f(n_2) : f(n_2,n_3)/f(n_3)$	70.49	63.11
$f(n_1,n_2)/f(n_2) : f(n_1,n_3)/f(n_3)$	80.32	66.39
$f(n_1,n_2) : f(n_2,n_3)$ (NEAR)	68.03	63.11
$f(n_1,n_2) : f(n_1,n_3)$ (NEAR)	71.31	67.21
$f(n_1,n_2)/f(n_1) : f(n_2,n_3)/f(n_2)$ (NEAR)	61.47	62.29
$f(n_1,n_2)/f(n_2) : f(n_2,n_3)/f(n_3)$ (NEAR)	65.57	57.37
$f(n_1,n_2)/f(n_2) : f(n_1,n_3)/f(n_3)$ (NEAR)	75.40	68.03#

Table 8: Performance of Altavista counts and BNC counts for compound bracketing (data from Lauer 1995)

Model	Accuracy
Baseline	63.93
Best BNC	68.03†‡
Lauer (1995): adjacency	68.90
Lauer (1995): dependency	77.50
Best Altavista	78.68†‡
Lauer (1995): tuned	80.70
Upper bound	81.50

Table 9: Performance comparison with the literature for compound bracketing

Noun Compound Bracketing (10)

(Nakov & Hearst 2005)

- Web page hits: proxy for *n*-gram frequencies
- Sample surface features
  - amino**-**acid sequence ➔ *left*
  - brain stem**'s** cell ➔ *left*
  - brain**'s** stem cell ➔ *right*
- Majority vote to combine the different models
- Accuracy **89.34%** (on the Lauer's set: baseline 66.70%, previous best result: 80.70%)

state of the art

193

Noun Compound Bracketing (11)

Problems with Web counts

- Page hits are inaccurate
  - maybe not that bad for same tasks? (cf. Keller & Lapata 2003)
- **The Web lacks linguistic annotation**
  - Pr(*health*| *care*) = #("*health care*") / #(*care*)
    - *health*: noun
    - *care*: both verb and noun
    - can be adjacent by chance
    - can come from different sentences
- Cannot find:
  - stem cells VERB PREPOSITION brain
  - protein synthesis' inhibition

194

Noun Compound Bracketing (12)

Solution proposed by (Nakov & Hearst 2005):

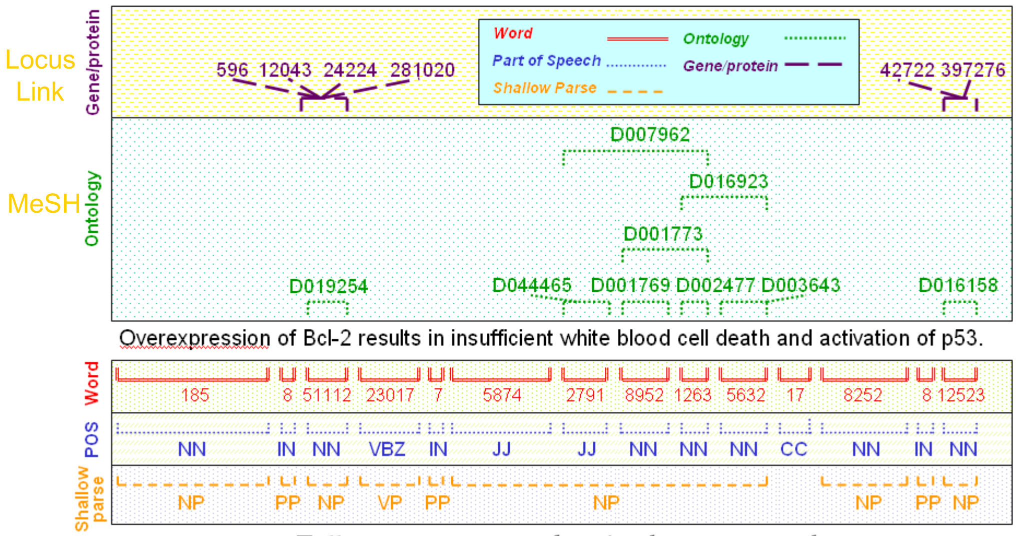
- MEDLINE: ~13M abstracts
  - they annotated:
    - ~1.4M abstracts
    - ~10M sentences
    - ~320M annotations
- Layered Query Language: http://biotext.berkeley.edu/lql/

Noun Compound Bracketing (13)

The System

- Built on top of an RDBMS system
- Supports layers of annotations over text
  - hierarchical, overlapping
  - cannot be represented by a single-file XML
- Specialized query language
  - LQL (Layered Query Language)

Noun Compound Bracketing (14)



***Full parse*, *sentence* and *section* layers are not shown.**

197

Noun Compound Bracketing (15)

Noun compound extraction

```
FROM
    [layer='shallow_parse' && tag_name='NP'
     ^ [layer='pos' && tag_name="noun"]
       [layer='pos' && tag_name="noun"]
       [layer='pos' && tag_name="noun"] $
    ] AS compound
SELECT compound.text
```

beginning of layer, but does not allow adj., det., etc.

By default: nothing can go in between

end of layer

198

Noun Compound Bracketing (16)

Finding bigram counts

```
SELECT COUNT(*) AS freq
FROM (
  BEGIN_LQL
    FROM
      [layer='shallow_parse' && tag_name='NP'
        [layer='pos' && tag_name="noun" &&
          content="immunodeficiency"] AS word1
        [layer='pos' && tag_name="noun" &&
          (content="virus"||content="viruses")]
      ]
    SELECT word1.content
  END_LQL
) AS word
ORDER BY freq DESC
```

just count

Inflections: UMLS lexicon

199

Noun Compound Bracketing (17)

Paraphrases

- Types of noun compounds (Warren,1978):
  - Paraphrasable
    - Prepositional
      - immunodeficiency virus **in** humans → *right*
    - Verbal
      - virus **causing** human immunodeficiency → left
      - immunodeficiency virus **found in** humans → *right*
    - Copula
      - immunodeficiency virus **that is** human → *right*
  - Other

200

Noun Compound Bracketing (18)

Prepositional paraphrases

```
SELECT LOWER(prp.content) lp, COUNT(*) AS freq
FROM (
  BEGIN_LQL
    FROM
      [layer='sentence'
         [layer='pos' && tag_name="noun" &&
            content = "immunodeficiency"]
         [layer='pos' && tag_name="noun" &&
            content IN ("virus","viruses")]
         [layer='pos' && tag_name='IN'] AS prep
        ?[layer='pos' && tag_name='DT' &&
            content IN ("the","a","an")]
         [layer='pos' && tag_name="noun" &&
            content IN ("human", "humans")]
       ] SELECT prep.content
   END_LQL
) AS prp
GROUP BY lp, ORDER BY freq DESC
```

optional layer

Noun Compound Bracketing (19)

Evaluation

- obtained 418,678 noun compounds (NCs)
- annotated the top 232 NCs
  - agreement 88%
  - kappa 0.606
- baseline (left): 83.19%
- *n*-grams: Pr, #, χ^2
- prepositional paraphrases

Noun Compound Bracketing (20)

Results

Model	$\surd$ (correct)	$\times$ (wrong)	$\emptyset$ (N/A)	**P**(%)	**C**(%)
# adjacency	196	36	0	84.48	100.00
Pr adjacency	173	59	0	74.57	100.00
χ^2 adjacency	200	32	0	86.21	100.00
# dependency	195	37	0	84.05	100.00
Pr dependency	193	39	0	83.19	100.00
χ^2 dependency	196	36	0	84.48	100.00
PrepPar	181	13	38	93.30	83.62
PP+χ^2adj+χ^2dep	207	13	12	94.09	94.83
PP+χ^2adj+χ^2dep$\rightarrow$right	214	18	0	**92.24**	100.00
Baseline (choose left)	193	39	0	83.19	100.00

203

Noun Compound Bracketing (21)

Discussion

- **Semantics** of bone marrow cells
  - top verbal paraphrases
    - cells **derived from** (the) bone marrow (22 instances)
    - cells **isolated from** (the) bone marrow (14 instances)
  - top prepositional paraphrases
    - cells **in** (the) bone marrow (456 instances)
    - cells **from** (the) bone marrow (108 instances)

cannot be down with/on the Web

- **Finding hard examples** for NC bracketing
  - $w_1 w_2 w_3$ such that both $w_1 w_2$ and $w_2 w_3$ are MeSH terms
- **Overall**:
  - Unsupervised approaches are fast and inexpensive, but they cannot capture all the cases, especially those that rely on meaning

204

Outline

205

Semantic Parsers for the Biology Domain (1)

Bioscience text processing:

Researchers (Hearst 2004) claim that bioscience semantics is both easier and harder than general, open-domain text:

- **Easier**: fewer ambiguities/subtleties;'systematic' meanings
- **Harder**: a large body of terminology; complex sentence structure

206

Semantic Parsers for the Biology Domain (2)

Example

"Recent research, in proliferating cells, has demonstrated that interaction of E2F1 with the p53 pathway could involve transcriptional up-regulation of E2F1 target genes such as p14/p19ARF, which affect p53 accumulation [67,68], E2F1-induced phosphorylation of p53 [69], or direct E2F1-p53 complex formation [70]."

(cf. Hearst 2004)

207

Semantic Parsers for the Biology Domain (3)

Semantic relations in Bioscience research

- specific relations, e.g.:
  - What is the role of this protein in that pathway?
  - Identify articles which show a direct proportional relationships between proteins/genes.
- There is need for:
  - Automatic discovery of semantic relations
    - Between nouns in noun compounds
    - Between entities in sentences
  - Acquisition of labeled data:
    - Idea: use text surrounding citations to documents to identify paraphrases

208

Semantic Parsers for the Biology Domain (4)

Discovering noun compound relations

- Technical (biomedical) text is rich with NCs
  - E.g.: Open-labeled long-term study of the subcutaneous sumatriptan efficacy and tolerability in acute migraine treatment.
- NC is any sequence of nouns that itself functions as a noun; E.g.:
  - migraine treatment
  - migraine treatment tolerability

209

Semantic Parsers for the Biology Domain (5)

Noun compound processing has three tasks:

- Identification
- Bracketing (syntactic analysis)
  - [baseline [headache frequency]]
  - [[tension headache] patient]
- **Semantic interpretation**
  - migraine treatment → treatment **for** headache
  - laser treatment → treatment **that uses** laser

210

Semantic Parsers for the Biology Domain (6)

How to interpret noun compounds?

- Idea:
  - Use the top levels of a lexical hierarchy to identify semantic relations (we've already seen this has been used in open-domain text as well)
- Hypothesis:
  - A particular semantic relation holds between all 2-word NCs that can be categorized by a lexical category pair.

211

Semantic Parsers for the Biology Domain (7)

One approach: Rosario and Hearst (2001)

- relations are pre-defined
- resources: MeSH, Neural Network
- 18 classification classes, 60% accuracy

212

Semantic Parsers for the Biology Domain (8)

The lexical hierarchy MeSH

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Physical Sciences [H]
9. Anthropology, Education, Sociology and Social Phenomena [I]
10. Technology and Food and Beverages [J]
11. Humanities [K]
12. Information Science [L]
13. Persons [M]
14. Health Care [N]
15. Geographic Locations [Z]

213

Semantic Parsers for the Biology Domain (9)

1. Anatomy [A] ⇨	Body Regions [A01]
2. [B]	Musculoskeletal System [A02]
3. [C]	Digestive System [A03]
4. [D]	Respiratory System [A04]
5. [E]	Urogenital System [A05]
6. [F]	……
7. [G]	
8. Physical Sciences [H]	
9. [I]	
10. [J]	
11. [K]	
12. [L]	
13. [M]	

214

Semantic Parsers for the Biology Domain (10)

1. Anatomy [A] ⇨	Body Regions [A01] ⇨	Abdomen [A01.047]
2. [B]	Musculoskeletal System [A02]	Back [A01.176]
3. [C]	Digestive System [A03]	Breast [A01.236]
4. [D]	Respiratory System [A04]	Extremities [A01.378]
5. [E]	Urogenital System [A05]	Head [A01.456]
6. [F]	……	Neck [A01.598]
7. [G]		….
8. Physical Sciences [H]		
9. [I]		
10. [J]		
11. [K]		
12. [L]		
13. [M]		

215

Semantic Parsers for the Biology Domain (11)

1. Anatomy [A] ⇨		Body Regions [A01] ⇨	Abdomen [A01.047]
2. [B]		Musculoskeletal System [A02]	Back [A01.176]
3. [C]		Digestive System [A03]	Breast [A01.236]
4. [D]		Respiratory System [A04]	Extremities [A01.378]
5. [E]		Urogenital System [A05]	Head [A01.456]
6. [F]		……	Neck [A01.598]
7. [G]			….
8. Physical Sciences [H] ⇨	Electronics		
9. [I]	Astronomy		
10. [J]	Nature		
11. [K]	Time		
12. [L]	Weights and Measures		
13. [M]	….		

216

Semantic Parsers for the Biology Domain (12)

1. Anatomy [A]	⇨	Body Regions [A01]	⇨	Abdomen [A01.047]
2. [B]		Musculoskeletal System [A02]		Back [A01.176]
3. [C]		Digestive System [A03]		Breast [A01.236]
4. [D]		Respiratory System [A04]		Extremities [A01.378]
5. [E]		Urogenital System [A05]		Head [A01.456]
6. [F]		......		Neck [A01.598]
7. [G]				....
8. Physical Sciences [H]	⇨	Electronics	⇨	Amplifiers
9. [I]		Astronomy		Electronics, Medical
10. [J]		Nature		Transducers
11. [K]		Time		
12. [L]		Weights and Measures		
13. [M]		....		

217

Semantic Parsers for the Biology Domain (13)

1. Anatomy [A]	⇨	Body Regions [A01]	⇨	Abdomen [A01.047]		
2. [B]		Musculoskeletal System [A02]		Back [A01.176]		
3. [C]		Digestive System [A03]		Breast [A01.236]		
4. [D]		Respiratory System [A04]		Extremities [A01.378]		
5. [E]		Urogenital System [A05]		Head [A01.456]		
6. [F]		......		Neck [A01.598]		
7. [G]				....		
8. Physical Sciences [H]	⇨	Electronics	⇨	Amplifiers		
9. [I]		Astronomy		Electronics, Medical		
10. [J]		Nature		Transducers		
11. [K]		Time				
12. [L]		Weights and Measures			⇨	Calibration
13. [M]		....				Metric System
						Reference Standard

218

Semantic Parsers for the Biology Domain (14)

Noun mapping to MeSH concepts

headache	pain
C23.888.592.612.441	G11.561.796.444

219

Semantic Parsers for the Biology Domain (15)

Noun mapping to MeSH concepts

headache	pain
C23.888.592.612.441	**G11**.561.796.444
↓	↓
Level 0	Level 0

220

Semantic Parsers for the Biology Domain (16)

Noun mapping to MeSH concepts

headache	pain
C23.888.592.612.441	**G11.561**.796.444
↓	↓
Level 1	Level 1

221

Semantic Parsers for the Biology Domain (17)

Noun mapping to MeSH concepts

headache	pain
C23.888.592.612.441	**G11.561.796**.444
↓	↓
Level 2	Level 2

..and so on

222

Semantic Parsers for the Biology Domain (18)

How does MeSH help?

- Idea:
  - Words in homogeneous MeSH subhierarchies behave "similarly" with respect to relation assignment
- Hypothesis:
  - A particular semantic relation holds between all 2-word NCs that can be categorized by a MeSH category pairs

223

Semantic Parsers for the Biology Domain (19)

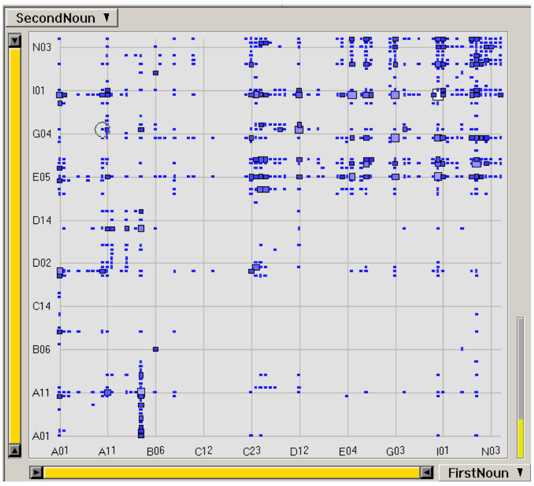
How to generalize noun compounds:
- CP = category pair

- CP: A02 C04 (Musculoskeletal System, Neoplasms)
  - skull tumors, bone cysts
- CP: C04 M01 (Neoplasms, Person)
  - leukemia survivor, lymphoma patients

224

Semantic Parsers for the Biology Domain (20)

Distribution of category pairs



225

Semantic Parsers for the Biology Domain (21)

Text collection:

- ~70,000 NCs extracted from titles and abstracts of Medline
- 2,627 CPs at level 0 (with at least 10 unique NCs)
  - analyzed
    - 250 CPs with Anatomy (A)
    - 21 CPs with Natural Science (H01)
    - 3 CPs with Neoplasm (C04)
  - This represents 10% of total CPs and 20% of total NCs

226

Semantic Parsers for the Biology Domain (22)

Classification algorithm

- For each CP
  - Divide its NCs into training-testing sets
  - For training inspect NCs by hand
    - Start from level 0 0
    - While NCs are not all similar
      - descend one level of the hierarchy
    - Repeat until all NCs for that CP are similar

227

Semantic Parsers for the Biology Domain (23)

- A02 C04
- B06 B06
- C04 M01
  - C04 M01.643
  - C04 M01.526
- A01 H01
  - A01 H01.770
  - A01 H01.671
    - A01 H01.671.538
    - A01 H01.671.868
- A01 M01
  - A01 M01.643
  - A01 M01.526
  - A01 M01.898

228

Semantic Parsers for the Biology Domain (24)

- A02 C04 → Location of Disease
- B06 B06 → Kind of Plants
- C04 M01
  - C04 M01.643 → Person afflicted by Disease
  - C04 M01.526 → Person who treats Disease
- A01 H01
  - A01 H01.770
  - A01 H01.671
    - A01 H01.671.538
    - A01 H01.671.868
- A01 M01
  - A01 M01.643
  - A01 M01.526
  - A01 M01.898

229

Semantic Parsers for the Biology Domain (25)

- A02 C04 → Location of Disease
- B06 B06 → Kind of Plants
- C04 M01
  - C04 M01.643 → Person afflicted by Disease
  - C04 M01.526 → Person who treats Disease
- A01 H01
  - A01 H01.770
  - A01 H01.671
    - A01 H01.671.538
    - A01 H01.671.868
- A01 M01
  - A01 M01.643 → Person afflicted by Disease
  - A01 M01.526
  - A01 M01.898

230

Semantic Parsers for the Biology Domain (26)

How many levels of specialization?

- Anatomy: 250 CPs
  - 187 (75%) remain first level
  - 56 (22%) descend one level
  - 7 (3%) descend two levels
- Natural Science (H01): 21 CPs
  - 1 (4%) remain first level
  - 8 (39%) descend one level
  - 12 (57%) descend two levels
- Neoplasms (C04) 3 CPs:
  - 3 (100%) descend one level

231

Semantic Parsers for the Biology Domain (27)

Evaluation:

- Test the decisions on "testing" set
- Count how many NCs that fall in the groups defined in the classification decisions are similar to each other
- Total Accuracy : 90.8%
- Generalization:
  - 415 classification decisions cover ~ 46,000 possible CP pairs

232

Semantic Parsers for the Biology Domain (28)

Challenges:

- **Lexical ambiguity**:
  - ***E.g.:*** mortality
    - state of being mortal
    - death rate
- **Relationship ambiguity** (rare cases):
  - ***E.g.:*** bacteria mortality
    - death *of* bacteria
    - death *caused by* bacteria

233

Semantic Parsers for the Biology Domain (29)

Conclusions

- Very simple method for assigning semantic relations to two-word technical NCs
  - Accuracy: 90.8%
- Lexical resource (MeSH) proves useful for this task
- Much less ambiguity in technical text vs. open-domain text; however, both this approach and Semantic Scattering prove that noun compounds need to be treated conceptually

234

Semantic Parsers for the Biology Domain (30)

Acquiring examples of semantic relations using citances

(Nakov, Schwartz, Hearst, SIGIR 2004)

- statements are backed up with a cite.
- papers are cited a lot (~30-100 times)
- Citances = the text around the citation which tends to state biological facts

- Different citances state the same facts in different ways, so they are used to create language models expressing semantic relations?

235

Semantic Parsers for the Biology Domain (31)

Why would we need citances?

- creation of training and testing data for semantic analysis,
- synonym set creation,
- document summarization,
- and information retrieval generally.

- Some preliminary results:
  - Citances are good candidates for paraphrase creation, and thus semantic interpretation.

236

Semantic Parsers for the Biology Domain (32)

How to do it?

- R(A, B) can be expressed in many ways:
  - R = a type of relation
  - A, B = types of entities
- Use citances to build a model which captures the different ways the relationship is expressed:
  - Seed learning algorithms with examples that
    - mention A and B,
    - R(A, B) holds.
  - Train a model to recognize R when the relation is not known.

237

Semantic Parsers for the Biology Domain (33)

- Identify a citance: appropriate phrase / clause / sentence that expresses it
- Grouping citances by topic
  - Citances that cite the same document should be grouped by the facts they state
- Normalize or paraphrase citances
  - Useful for
    - IR,
    - Text summarization,
    - Learning synonyms,
    - Relation extraction,
    - Question answering, and
    - Machine translation

238

Semantic Parsers for the Biology Domain (34)

Previous work

- Citation analysis goes back to the 1960's and includes:
  - Citation categorization,
  - Context analysis,
  - Citer motivation
- Citation indexing systems (e.g.: ISI's SCI, and CiteSeer)

239

Semantic Parsers for the Biology Domain (35)

Examples

- *NGF* withdrawal from sympathetic neurons induces *Bim*, which then contributes to death.
- *Nerve growth factor* withdrawal induces the expression of *Bim* and mediates Bax dependent cytochrome c release and apoptosis.
- The proapoptotic Bcl-2 family member *Bim* is strongly induced in sympathetic neurons in response to *NGF* withdrawal.
- In neurons, the BH3 only Bcl2 member, *Bim*, and JNK are both implicated in apoptosis caused by *nerve growth factor* deprivation.

240

Semantic Parsers for the Biology Domain (36)

Paraphrases of these examples:

- NGF withdrawal induces Bim.
- Nerve growth factor withdrawal induces the expression of Bim.
- Bim has been shown to be upregulated following nerve growth factor withdrawal.
- Bim implicated in apoptosis caused by nerve growth factor deprivation.

They all paraphrase:

Bim is induced after NGF withdrawal.

241

Semantic Parsers for the Biology Domain (37)

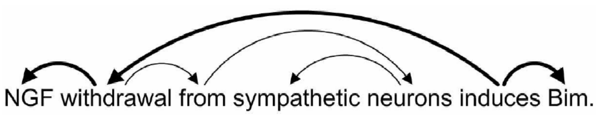
Algorithm of Paraphrase Creation

1. Extract the sentences that cite the target
2. Mark the named entities (NEs) of interest (genes/proteins, MeSH terms) and normalize.
3. Dependency parse (e.g.: MiniPar)
4. For each parse

   For each pair of NEs of interest

   i. Extract the path between them.

   ii. Create a paraphrase from the path.
5. Rank the candidates for a given pair of NEs.
6. Select only the ones above a threshold.
7. Generalize.



242

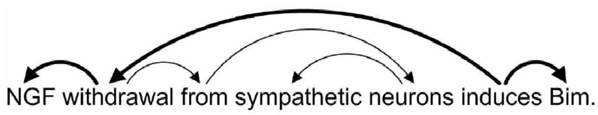
Semantic Parsers for the Biology Domain (38)

Given the path from the dependency parse:

Restore the original word order.

Add words to improve grammaticality.

- Bim … shown … be … following nerve growth factor withdrawal.
- Bim [has] [been] shown [to] be [upregulated] following nerve growth factor withdrawal.



243

Semantic Parsers for the Biology Domain (39)

Examples:

- NGF withdrawal induces Bim.
- Nerve growth factor withdrawal induces [the] expression of Bim.
- Bim [has] [been] shown [to] be [upregulated] following nerve growth factor withdrawal.
- Bim [is] induced in [sympathetic] neurons in response to NGF withdrawal.
- member Bim implicated in apoptosis caused by nerve growth factor deprivation.

244

Semantic Parsers for the Biology Domain (40)

System evaluation

- An influential journal paper from *Neuron*:
  - J. Whitfield, S. Neame, L. Paquet, O. Bernard, and J. Ham. Dominantnegative c-jun promotes neuronal survival by reducing bim expression and inhibiting mitochondrial cytochrome c release. Neuron, 29:629–643, 2001.
- 99 journal papers citing it
- 203 citances in total
- 36 different types of important biological factoids
  - But we concentrated on one model sentence:

    "Bim is induced after NGF withdrawal."

245

Semantic Parsers for the Biology Domain (41)

- **Set 1:**
  - 67 citances pointing to the target paper and manually found to contain a good or acceptable paraphrase (do not necessarily contain *Bim* or *NGF*);
- **Set 2:**
  - 65 citances pointing to the target paper and containing both *Bim* and *NGF*;
- **Set 3:**
  - 102 sentences from the 99 texts, containing both *Bim* and *NGF*

246

Semantic Parsers for the Biology Domain (42)

Correctness evaluation:

- *Bad* (0.0), if:
  - different relation (often phosphorylation aspect);
  - opposite meaning;
  - vagueness (wording not clear enough).
- *Acceptable* (0.5), If it was not *Bad* and:
  - contains additional terms (e.g., DP5 protein) or topics (e.g., PPs like in sympathetic neurons);
  - the relation was suggested but not definitely.
- Else *Good* (1.0)

247

Semantic Parsers for the Biology Domain (43)

Results:

- **Obtained 55, 65 and 102 paraphrases for sets 1, 2 and 3**
- **Only one paraphrase from each sentence**

  comparison of the dependency path to that of the model sentence

	correctness				grammaticality			
set	**1.0**	**0.5**	**0.0**	**%**	**1.0**	**0.5**	**0.0**	**%**
1	20	25	10	81.82	22	19	14	74.55
2	20	25	20	69.23	26	22	17	73.58
3	25	38	39	61.76	42	22	38	62.75
cluster	16	15	11	73.81	20	12	10	76.19

% - good (1.0) or acceptable (0.5)

248

Semantic Parsers for the Biology Domain (44)

Correctness (Recall)

- Calculated on Set 1
- 60 paraphrases (out of 67 citances)
- 5 citances produced 2 paraphrases
- *system recall:* 55/67, i.e. 82.09%
- 10 of the 67 relevant in Set 1 initially missed by the human annotator
  - 8 good,
  - 2 acceptable.
- *human recall* is 57/67, i.e. 85.07%

249

Semantic Parsers for the Biology Domain (45)

Grammaticality

- Missing coordinating "*and*":
  - "Hrk/DP5 Bim [have] [been] found [to] be upregulated after NGF withdrawal"
- Verb subcategorization
  - "caused by NGF role for Bim"
- Extra subject words
  - member Bim implicated in apoptosis caused by NGF deprivation
  - *sentence:* "In neurons, the BH3-only Bcl2 member, Bim, and JNK are both implicated in apoptosis caused by NGF deprivation."

250

Outline

251

KB Construction (1)

Knowledge Acquisition

The Problem:

- Building knowledge bases (i.e., ontologies) by hand is a laborious, time-intensive process that requires domain expertise (and possibly, knowledge representation expertise).

The Solution:

- Automatic extraction of important concepts and semantic relations from text documents.
- Advantages:
  - less domain expertise required up front
  - less time required to acquire knowledge for a new domain
  - provides greater flexibility than traditional approaches

252

KB Construction (2)

- Ontology definition
  - a set of concepts and semantic relations (vocabulary and facts)
- "Upper" ontologies
  - Embodies general purpose knowledge
    - Includes facts that an average person would take for granted.

253

KB Construction (3)

- Examples of ontologies:
  - WordNet – a publicly available lexical database, also contains some semantic relations. This information has been the most readily available to developers
  - IEEE's SUMO – standard Upper Merged Ontology
  - Others (CYC/OpenCYC)
- Domain-specific ontologies
  - Used to "extend" the general purpose, upper ontologies.

254

KB Construction (4)

KAT- Basic Idea

(Moldovan & Girju 2001)

- Example:
  "When the US economy enters a boom,
  mortgage interest rates rise."

(1) new concept *mortgage interest rate*

(2) state of *US economy* and the value of mortgage interest rate are in a DIRECT RELATIONSHIP

255

KB Construction (5)

Ontology Overview

- Supports knowledge-based reasoning:
  - Assume concepts labeled A-H
  - These concepts are linked via "is-a" and "has" relations
  - Given concept D, can we determine if D has F,C, and/or H?

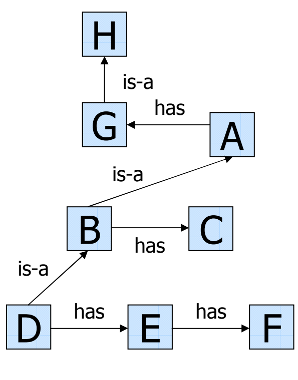
  Solution: navigate the graph:

  F: D has E, E has F; D has F

  C: D is B, B has C; D has C

  H: D is B, B is A, A has G,

  G is H; D has H.



256

KB Construction (6)

KAT Approach

Steps taken to acquire new knowledge:

- Identify seed concepts
- Build corpus of sentences related to (containing) seed concepts
- Analyze corpus for new concepts and relations
- Classify resulting concepts
- Link results with knowledge base

257

KB Construction (7)

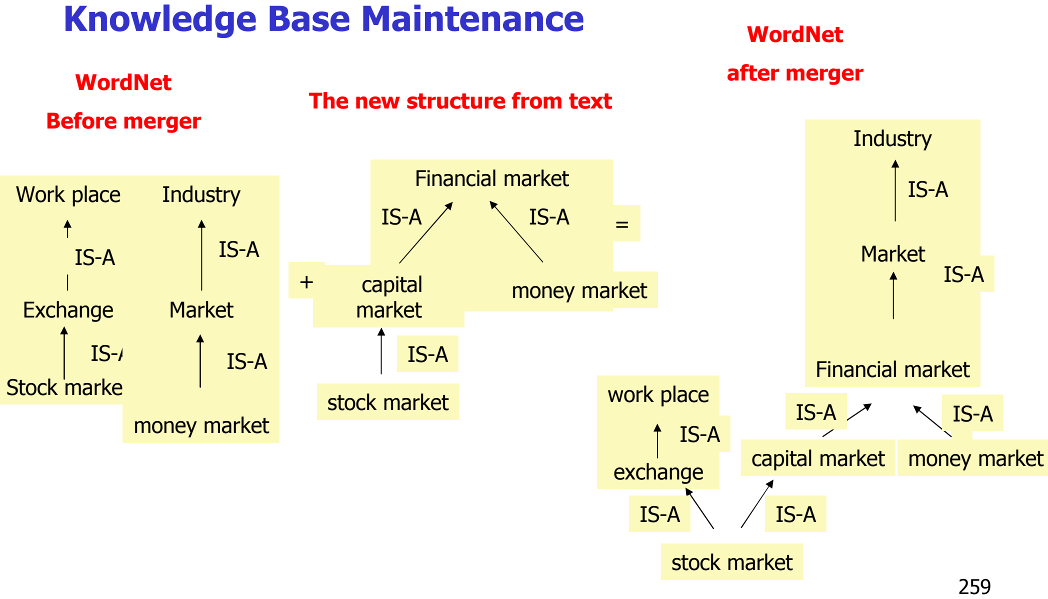
Classification – Overview

- After gathering concepts and relations, we now must classify the results in the ontology
- Goal of classification:
  - establishment of semantic "links" between discovered concepts and concepts in a known, general purpose knowledge base (i.e., WordNet)
- Specifically, provide an IS-A link to "connect" the hypernymy tree
  - The concepts for which we discover (in the corpus) IS-A relations, this is likely not a huge task
  - the main challenge is to decide how to establish an IS-A classification for those concepts LACKING a directly discovered IS-A relation
  - general approach focuses on the head word as the starting point

258

KB Construction (8)

Knowledge Base Maintenance



259

KB Construction (9)

Military Results

Documents examined	4259
Sentences parsed	143971
Sentences retained	31781
Retained concepts	44872
Discovered seed concepts	4689
Discovered non-seed concepts	1067
Concept precision (estimate)	87%
IS-A relations detected (relevant)	290
IS-A precision, total (estimate)	20%
IS-A precision, only relevant (estimate)	60%

260

KB Construction (10)

Military results – new "aircraft" terms

heavier-than-air craft #1
- airplane #1
  - bomber #1
    - strategic bomber #1
  - fighter #2
    - interceptor #1
      - missile interceptor #1
        - us missile interceptor #1
        - united states missile interceptor #1
        - air defense missile interceptor #1
  - jet #1
    - cargo jet #1
      - boeing 747 cargo jet #1
        - air boeing 747 cargo jet #1
          - iran air boeing 747 cargo jet #1
    - fighter jets #1
      - israeli fighter jets #1
- helicopter #1
  - anti-submarine helicopter #1

261

KB Construction (11)

Terrorism Results

Documents examined	1178
Sentences parsed	50115
Sentences retained	1768
Retained concepts	2107
Discovered seed concepts	592
Discovered non-seed concepts	137
Concept precision (estimate)	80
IS-A relations detected (relevant)	191
IS-A precision, total (estimate)	20%
IS-A precision, only relevant	58%

262

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Applications of semantic relations
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

263

Question Answering (1)

- A QA system:
  - Input: a natural language question as input
  - Output: a concise answer to the question
- A QA system is not a search engine
- Example questions:
  - What was the monetary value of the Nobel Peace Prize in 1989?
  - What does the Peugeot company manufacture?
  - How did Socrates die?

264

Question Answering (2)

QA systems today:

- QA systems have to answer a set of ***factoid questions*** (~500)
- In the first three years, the systems were asked to return 5 ranked text passages (50/250 bytes) to each question:
  - **Mean Reciprocal Rank** scoring:
    - 1, 0.5, 0.33, 0.25, 0.2, 0 (for 1st, 2nd, 3rd, 4th, 5th, rest);
- Starting 2002, a single exact answer was required based on the notion of confidence;

265

Question Answering (3)

Q: When did Lincoln die?

A:

1. During the civil war
2. In the spring time
3. at a theatre
4. April 15, 1865 ***
5. In April; 1965

MRR =1/4 = 0.25

Ok. So where do semantic relations fit in?

266

Question Answering (4)

Knowledge intensive applications

E.g. ***Question Answering***

Q: What does the [BMW company]IS-A produce?
A: "[BMW cars]MAKE-PRODUCE are sold .."

Q: Where have nuclear incidents occurred?
A: "The [(Three Mile Island) (nuclear incident)]LOC caused a DOE policy crisis.."

Q: What causes malaria?
A: "..to protect themselves and others from being bitten by [malaria mosquitoes]CAUSE.."

267

Question Answering (5)

Knowledge intensive applications

Q: What does the AH-64A Apache helicopter consist of?

(Girju et al., 2003)

A:
AH-64A Apache helicopter
Hellfire air to surface missille
millimeter wave seeker
70mm Folding Fin Aerial rocket
30mm Cannon camera
armaments
General Electric 1700-GE engine
4-rail launchers
four-bladed main rotor
anti-tank laser guided missile
Longbow millimetre wave fire control radar
integrated radar frequency interferometer
rotating turret
tandem cockpit
Kevlar seats

(Defense Industries: www.army-technology.com)

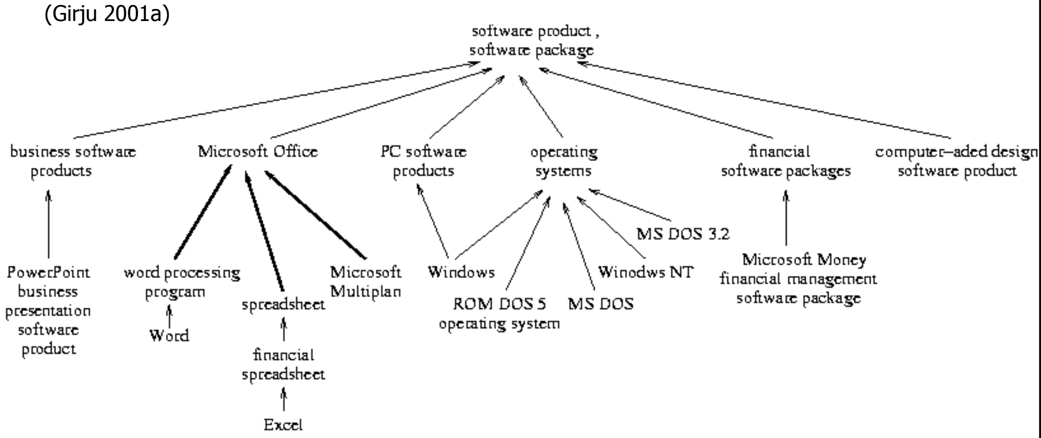
268

Question Answering (6)

Knowledge intensive applications

Q: What *software products* does Microsoft sell?

(Girju 2001a)



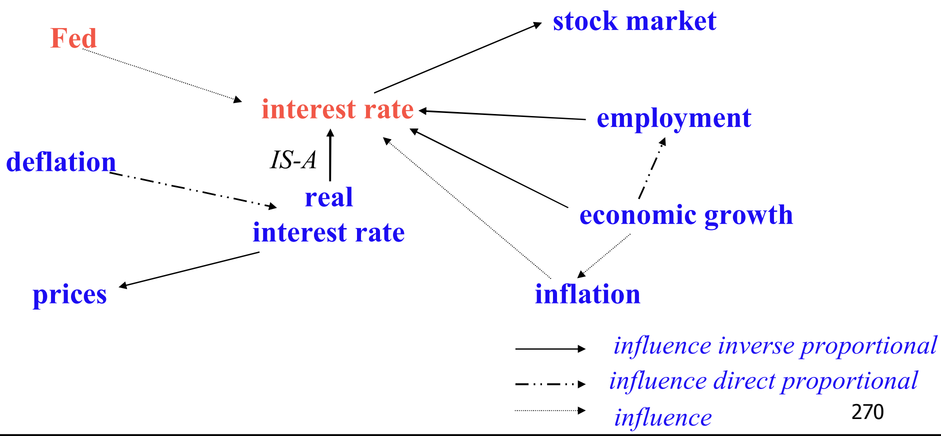
269

Question Answering (7)

Knowledge intensive applications

Q: Will the Fed change interest rate at their next meeting?

(Girju 2001b)



270

Outline

1. Introduction
   - The problem of knowledge discovery
   - Motivation
   - Basic approaches
   - Semantic relation discovery- The challenges
2. Lists of semantic relations
   - Approaches in Linguistics
   - Approaches in Natural Language Processing
3. Architectures of semantic parsers
   - Paraphrasing / Similarity-based systems
   - Conceptual-based systems
   - Context-based / hybrid systems – SemEval 2007, Task4
4. Going beyond base-NPs: the task of noun compound bracketing
5. Semantic parsers for the biology domain
6. Applications of semantic relations
   - KB construction
   - Question answering
   - Textual Entailment
   - Text-to-Scene Generation
7. Future Trends
8. Bibliography

Textual Entailment (1)

The PASCAL Semantic Entailment Task (Fall 2004 – current)

- T: "Chretien visited Peugeot's newly renovated car factory".
- H: "Peugeot manufactures cars".

T => H?

Textual Entailment (2)

What about semantic relations?

- T: "Chretien visited Peugeot's newly renovated **car factory**".
- H: "Peugeot manufactures cars".

T => H?

273

Textual Entailment (3)

Semantic relation detection for Textual Entailment

- Monotonicity of semantic relations
  - In compositional semantics, meanings are seen as functions, and can have various *monotonicity* properties:

    Upward monotone

    Downward monotone

274

Textual Entailment (4)

Upward-monotone (↑M)
The default: from small to large

Example: *broken*. Since *chair* IS-A *furniture*, *broken chair* => *broken furniture*
Heuristic: in a ↑M context, broadening context preserve truth

Downward-monotone (↓M)
Negatives, restrictives, etc.: from big to small

Example: *doesn't*. While *hover* IS-A *fly*, *doesn't fly* => *doesn't hover*
Heuristic: in a ↓M context, narrowing context preserve truth

Textual Entailment (5)

- Monotonicity is imposed by quantifiers (determiners) D:
  - Each determiner takes two sets as arguments — one corresponding to its ***restriction (A)***, and the other corresponding to its ***nuclear scope (B)***.
- We may evaluate the restriction and the nuclear scope as separate "environments" in which negative polarity items might occur.

Textual Entailment (6)

Definition

We say a determiner **D creates an upward entailing environment in its restriction** iff the following condition holds (for all A, B, C):

$$[\ [D(A,B)\ \&\ A \subseteq C] \rightarrow D(C,B)\ ]$$

277

Textual Entailment (7)

Definition

We say a determiner **D creates a downward entailing environment in its restriction** iff the following condition holds (for all A, B, C):

$$[\ [D(A,B)\ \&\ C \subseteq A] \rightarrow D(C,B)\ ]$$

278

Textual Entailment (8)

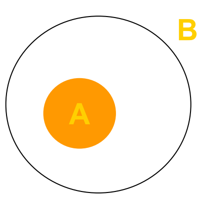
How do we following determiners according to the environments they create in their restrictions and scopes:

- SOME(?,?) Some dogs bark.
- NO(?,?) No dog barks.
- EVERY(?,?) Every dog barks.

Textual Entailment (9)

Example:

- EVERY(?,?) EVERY(A, B): $A \subseteq B$



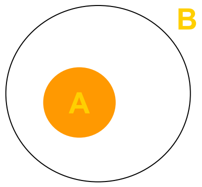
Upward: [[D(A,B) & $A \subseteq C$] $\rightarrow$ D(C,B)]
Downward: [[D(A,B) & $C \subseteq A$] $\rightarrow$ D(C,B)]

Upward or Downward in its restriction?
(Left Upward or Left Downward?)

E.g.: Every duck flies.

Textual Entailment (10)

- EVERY(?,?) EVERY(A, B): $A \subseteq B$

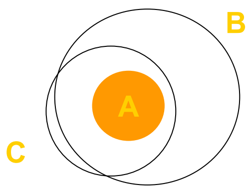


Upward: [[D(A,B) & $A \subseteq$ **C**] $\rightarrow$ D(C,B)]

E.g.: Every duck flies.

Textual Entailment (11)

- EVERY(?,?) EVERY(A, B): $A \subseteq B$



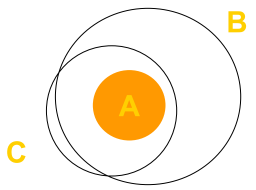
Upward: [[D(A,B) & $A \subseteq$ **C**] $\rightarrow$ D(C,B)]

E.g.: Every duck flies. $\overset{?}{=>}$ Every bird flies.

duck IS-A bird

Textual Entailment (12)

- EVERY(?,?) EVERY(A, B): $A \subseteq B$



<u>Upward</u>: $[\ [\cancel{D(A,B)\ \&\ A \subseteq \mathbf{C}}] \rightarrow D(C,B)\ ]$

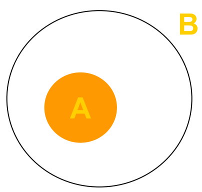
E.g.: Every duck flies. => Every bird flies.

duck IS-A bird

283

Textual Entailment (13)

- EVERY(?,?) EVERY(A, B): $A \subseteq B$



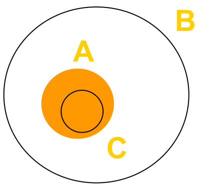
<u>Downward</u>: $[\ [D(A,B)\ \&\ C \subseteq A] \rightarrow D(C,B)\ ]$

E.g.: Every duck flies.

284

Textual Entailment (14)

- EVERY(?,?) EVERY(A, B): $A \subseteq B$



<u>Downward</u>: [[D(A,B) & $C \subseteq A$] $\rightarrow$ D(C,B)]

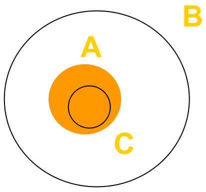
E.g.: Every duck flies. $\overset{?}{=>}$ Every mallard flies.

mallard IS-A duck

285

Textual Entailment (15)

- EVERY($\Downarrow$,?) EVERY(A, B): $A \subseteq B$



<u>Downward</u>: [[D(A,B) & $C \subseteq A$] $\rightarrow$ D(C,B)]

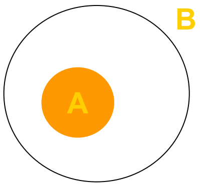
E.g.: Every duck flies. => Every mallard flies. ✓

mallard IS-A duck

286

Textual Entailment (16)

- EVERY(⇓,?) EVERY(A, B): $A \subseteq B$



<u>Upward</u>: [[D(A,B) & $B \subseteq C$] → D(A,C)]
<u>Downward</u>: [[D(A,B) & $C \subseteq B$] → D(A,C)]

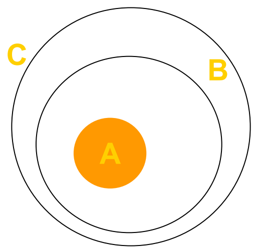
Upward or Downward in its nuclear scope?
(Upward Right or Downward Right?)

E.g.: Every duck flies.

287

Textual Entailment (17)

- EVERY(⇓,?) EVERY(A, B): $A \subseteq B$



<u>Upward</u>: [[D(A,B) & $B \subseteq$ **C**] → D(A,C)]

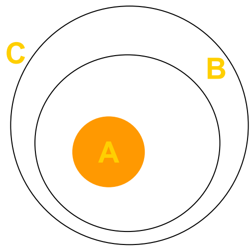
E.g.: Every duck flies. $\overset{?}{=>}$ Every duck moves.

flies IS-A moves

288

Textual Entailment (18)

- EVERY(⇓,?) EVERY(A, B): A ⊆ B



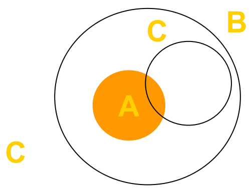
Upward: [[D(A,B) & B ⊆ **C**] → D(A,C)]

E.g.: Every duck flies. => Every duck moves. ✓

flies IS-A moves

Textual Entailment (19)

- EVERY(⇓,?) EVERY(A, B): A ⊆ B



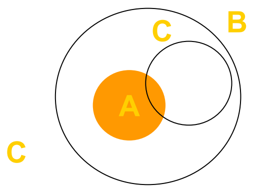
Downward: [[D(A,B) & C ⊆ B] → D(A,C)]

E.g.: Every duck flies. ?=> Every duck hovers.

flies IS-A hovers

Textual Entailment (20)

- **EVERY(⇓, ⇑)**   EVERY(A, B): A $\subseteq$ B



E.g.: Every duck flies. $\not\Rightarrow$ Every duck hovers.

flies IS-A hovers

291

Textual Entailment (21)

- What about other relations besides IS-A?
  - Cause-Effect
  - Part-Whole
  - Etc.

- This is left for future research

292

Outline

293

Text-to-Scene Generation (1)

- Idea: Give a snipet of text, generate an image that is a faithful representation of that text
- Application:
  - Education
  - To help in insurance litigations
  - For theoretical reasons
- Challenges:
  - not every piece of text can receive a pictorial representation (e.g. abstract words such as policy, government, feeling, etc.)
  - Some knowledge need to represent pictures is not explicitly stated in text, but inferred
- State of the art systems:
  - WordsEye (Coyle and Sproat 2001)
  - CarSim ()

294

Text-to-Scene Generation (2)

WordsEye (Coyle and Sproat 2001)

- impressive system that recreates 3D animated scenes from short descriptions.
- The number of 3D objects WordsEye uses - 12,000 - gives an idea of its ambition.
- WordsEye integrates state-of-the-art resources
  - Collins' dependency parser
  - the WordNet lexical database

295

Text-to-Scene Generation (3)

WordsEye (Coyle and Sproat 2001)

- The interpretation of a narrative is based on an extension of case grammars (semantic frames) and a good deal of inferences about the environment (Sproat 2001).
- **Accepts sentences that** describe the positions of common objects.
  - E.g.,The cat is on the table
- Gradually build scene by adding objects, colors, textures, sizes, orientations ...

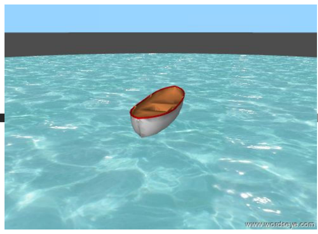
  E.g., The cat is on the **large** chair.
  A dog is **facing** the chair.
  A **brick** wall is **2 feet behind** the dog.
  The wall is **20 feet wide**.
  The ground is **pale green**.

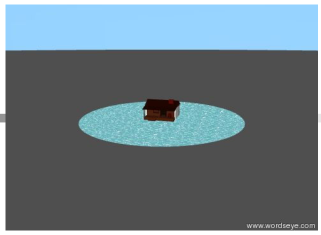
296

Text-to-Scene Generation (4)

- Limitations:
  - No verbs/poses (eg, *running, happy* ...)
  - Doesn't do well on object parts (eg, *hood of the car*)
  - Does not address real world stories.
  - The example narratives resemble imaginary fairy tales.

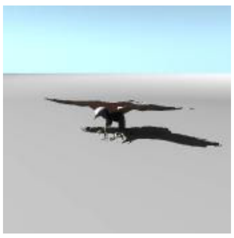
http://www.wordseye.com/





boat on the lake

vs. cabin on the lake



eagle in the nest

vs. eagle in the sky

flowers in a vase

Text-to-Scene Generation (6)

CarSim: Text-to-Scene Conversion for Accident Visualization

- a system that automatically converts textual descriptions of accidents into 3D scenes.
- combines language processing and visualization techniques.
- Input: a written report describing an accident as input.
- A first module analyzes the report using information extraction techniques and produces a representation of it.

Text-to-Scene Generation (7)

CarSim: Text-to-Scene Conversion for Accident Visualization

- A visualization module constructs a 3D scene from it and replays the accident symbolically.
- prototypes in French and English, but only the Swedish version of Carsim is available online.
- CarSim has been applied to a corpus of 87 reports written in French for which it can currently synthesize visually 35 percent of the texts
  - These texts are real reports collected from an insurance company.
  - They have been written by drivers after their accidents and transcribed without modification.

Text-to-Scene Generation (8)

Example of input text:

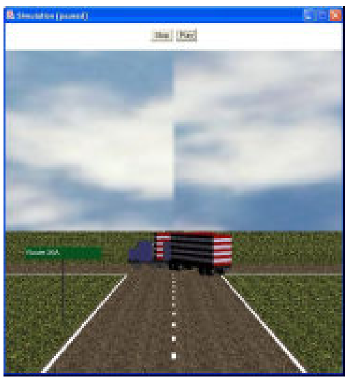
Je roulais sur la partie droite de la chaussée quand un véhicule arrivant en face dans le virage a été complètement déporté. Serrant à droite au maximum, je n'ai pu éviter la voiture qui arrivait à grande vitesse. (Report A8, MAIF corpus)

I was driving on the right-hand side of the road when a vehicle coming in front of me in the bend skidded completely. Moving to the right of the lane as far as I could; I couldn't avoid the car that was coming very fast. (author's translation)

301

Text-to-Scene Generation (9)

Examples of scenes generated



302

Outline

303

References

- Baker C., Fillmore C., and Lowe J. 1998. *The Berkeley FrameNet Project*.
- Barker K. and Szpakowicz S 1998. Semi-automatic recognition of noun modifier relationships. In *Proc. of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 96-102, Montreal, Canada.
- Blaheta D. and Charniak E. *Assigning function tags to parsed text*. In Proc. NAACL-00, 2000.
- Brill E. 1997. Natural Language Processing Using Very Large Corpora. Kluwer Academic Press.1997.
- Carlson, G. (1984) On the Role of Thematic Roles in Linguistic Theory. Linguistics, 22, 259-279.
- Charniak E. 2000. A Maximum-Entropy-Inspired Parser In Proc. of the Conference of the
- North American Chapter of the Association for Computational Linguistics (NAACL), pages 132-139.
- Chomsky, N. (1986) *Knowledge of Language*: *Its Nature, Origin and Use*. New York: Praeger.
- Croft, W.(1991). *Syntactic Categories and Grammatical Relations*. Univ. of Chicago Press.
- Cruse, D. A.1973. Some Thoughts on Agentivity. *Journal of Linguistics* 9: 1-204.

304

References

- Collins M. 1997. Three Generative, Lexicalised Models for Statistical Parsing. Proceedings of the 35th Annual Meeting of the ACL (jointly with the 8th Conference of the EACL), Madrid.
- Dik, S.C. (1989). *The Theory of Functional Grammar, Part I: The Structure of the Clause*, Foris Publications, Dordrecht.
- Downing, P. 1977. On the creation and use of English compound nouns. *Language* 53, pp. 810-842.
- Dowty, D. R. (1979). *Word Meaning and Montague Grammar*, Kluver Academic Publishers.
- Dowty, D., On the Semantic Content of the Notion of Thematic Role, in G. Cherchia, B. Partee, R. Turner (eds), *Properties, Types and meaning*, Kluwer, 1989.
- Dowty, D., Thematic Proto-roles and Argument Selection, Language, vol. 67-3, 1991.
- Fillmore, C., The Case for Case Reopened. In P. Cole and J. Sadock *Syntax and Semantics 8: Grammatical Relations*, Academic Press, New York, pp. 59-82 1977.
- Fillmore, C., The Case for Case, in *Universals in Linguistic Theory*, E. Bach and R.T. Hams (eds.), Holt, Rinehart and Winston, New York, 1968.
- Fillmore, C. and B.T.S. Atkins. 2000. Describing Polysemy: The Case of 'Crawl'. In Ravin ed. Polysemy: Theoretical and Computational Approaches.91-110. Oxford, Oxford University Press.

305

References

- Fillmore, C. and Collin F. Baker (2001): Frame Semantics for Text Understanding. In Proceedings of WordNet and Other Lexical Resources Workshop, NAACL, Pittsburgh, June, 2001
- Finin, T.W. 1980. The semantic interpretation of compound nominals. University of Illinois at Urbana-Champaign. University Microfilms International.
- Girju R. 2001. Answer Fusion with On-Line Ontology Development . In Proceedings of the *North American Chapter of the Association for Computational Linguistics (NAACL) - Student Research Workshop*, (NAACL 2001), Pittsburgh, PA, June 2001.
- Girju R., Badulescu A., and Moldovan D. 2003. Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations. In the Proceedings of the *Human Language Technology Conference*, Edmonton, Canada, May-June 2003.
- Gruber, J. (1967) *Studies in Lexical Relations*, MIT doctoral dissertation and in *Lexical Structures in Syntrax and Semantics*, North Holland (1976).
- Isabelle, P. 1984. "Another look at nominal compounds" in *Proceedings of the 22nd Annual ACL Conference, COLING-84*, Stanford, CA., pp. 509-516.
- Jackendoff, R. 1990. Semantic Structures, Cambridge, Mass., The MIT Press.
- Jackendoff, R., *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge, 1972.

306

References

- Jespersen, O. 1954. *A modern English grammar on historical principles*, VI. (George Allen & Unwin Ltd., London, 1909-49; reprinted 1954).
- *Kingsbury P., Palmer P., and Marcus M. 2002*. Adding Semantic Annotation to the Penn TreeBank. In Proceedings of the Human Language Technology Conference, San Diego, California.
- *Kingsbury P. and Palmer P. 2002.* From Treebank to PropBank *In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain.*
- Lakoff, George.1977. Linguistic Gestalts. *Chicago Linguistics Society* 13: 236-87.
- Lapata M. 2000. The Automatic Interpretation of Nominalizations. In Proceedings of AAAI, pages 716-721
- Lauer, M. 1994. "Conceptual Association for Compound Noun Analysis" in *Proceedings of the 32nd Annual ACL Conference*, Las Cruces, N.M., pp.474-481.
- Lees, R.B. 1960. *The grammar of English nominalizations*. Indiana University, Bloomington, Indiana. Fourth printing 1966.
- Lees, R.B. 1970. "Problems in the grammatical analysis of English nominal compounds" in *Progress in Linguistics*, eds. Bierwisch, M., and K.E. Heidolph, The Hague, Mouton, pp. 174-186.

References

- Leonard, R. 1984. *The interpretation of English noun sequences on the computer*. North-Holland Linguistic Studies, Elsevier, the Netherlands.
- Levi, J.N. 1978. *The syntax and semantics of complex nominals*. Academic Press, New York.
- Levin B. and Rappaport-Hovav M.1996. From Lexical Semantics to Argument Realization. unpublished ms., Northwestern University and Bar Ilan University, Evanston, IL and Ramat Gan, Israel (http://www.ling.nwu.edu/~beth/pubs.html)
- Liberman M. and Church K.W. Text analysis and word pronunciation in text-to-speech synthesis. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*. Dekker, 1991.
- Lin D. and Patel P. 2002. Concept Discovery from Text. In *Proceedings of Conference on Computational Linguistics 2002*. pp. 577-583. Taipei, Taiwan.
- McDonald, D. 1981. "Compound: a program that understands noun compounds" in *IJCAI-81, International Joint Conference on Artificial Intelligence*, p 1061.
- Miller, G. A. and Fellbaum, C. (1998), *Wordnet: An Electronic Lexical Database*, MIT Press**.**
- Moldovan D. and Girju R. 2001. An Interactive Tool For The Rapid Development of Knowledge Bases. In the *International Journal on Artificial Intelligence Tools (IJAIT)*, vol 10., no. 1-2, March 2001.

References

- Pollard C. and Sag I. 1994. Head-driven Phrase Structure Grammar. University Press, Chicago.
- Rappaport, M., Levin, B., What to do with -roles ?, in *Syntax and Semantics 21: Thematic Relations*, W. Wilkins (ed.), Academic Press, 1988.
- Riloff, E. 1993. *Automatically Constructing a Dictionary for Information Extraction Tasks*. In Proceedings of the Eleventh National Conference on Artificial Intelligence. 811--816.
- Riloff, E., and Schmelzenbach, M. 1998. *An Empirical Approach to Conceptual Case Frame Acquisition*. In Proceedings of the Sixth Workshop on Very Large Corpora.
- Rosario B. and Hearst M. 2001. "Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy". In the Proceedings of the 2001 Conference on EMNLP, pages 82-90.
- Rosario B., Hearst M., and Fillmore C. 2002. "The Descent of Hierarchy, and Selection in Relational Semantics". In the Proceedings of the Association for Computational Linguistics (ACL) 2002.
- TER STAL, WILCO (1996). "Automated Interpretation of Nominal Compounds in a Technical Domain." Ph.D. thesis, University of Twente, The Netherlands.

309

References

- Payne. T. 1997. Describing morphosyntax: A guide for field linguists. xvii, 413 pp. Cambridge: Cambridge University Press.
- van Valin, Robert D.1990. Semantic Parameters of Split Transitivity. *Language* 66: 221-60.
- Selkirk, E.O. 1982. *The syntax of words*. MIT Press, Cambridge, MA.
- Vanderwende, L. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of COLING-94*, Kyoto, Japan, pp. 782-788.
- W. ter Stal and P. van der Vet. Two-level semantic analysis of compounds: A case study in linguistic engineering. In *Papers from the 4th CLIN Meeting*, 1994.
- Williams, E. (1981a) ``Argument Structure and Morphology." *Linguistic Review*, 1, 81-114.

310

Thank You

311